



**This electronic thesis or dissertation has been  
downloaded from Explore Bristol Research,  
<http://research-information.bristol.ac.uk>**

*Author:*

**Zhang, Jinglu**

*Title:*

**Assessing and understanding Chinese high school students' scientific argumentation competence**

**General rights**

Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>. This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

**Take down policy**

Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact [collections-metadata@bristol.ac.uk](mailto:collections-metadata@bristol.ac.uk) and include the following information in your message:

- Your contact details
- Bibliographic details for the item, including a URL
- An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

# **Assessing and Understanding Chinese High School Students' Scientific Argumentation Competence**

Jinglu Zhang

School of Education, University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements for  
award of the degree of Doctor of Philosophy in the Faculty of Social Sciences and Law

Wordcount: 84312

## Abstract

Argumentation is an important practice in science by which knowledge is constructed, evaluated, and modified. Scientific argumentation (SA) is thus a promising activity in science education to enhance students' understanding of science. However, internationally, the lack of consensus on its nature, especially in the educational context, has led to an under-exploration of its assessment, which limits its integration into the science classroom. In the context of this study, China, there is a dilemma regarding SA. On the one hand, curriculum policy calls for equipping Chinese students with key competences needed in the future world. Accordingly, the high school Physics Curriculum targets SA as a key competence. However, on the other hand, due to the notoriously examination driven approach to education SA tends to be neglected since it is currently not assessed in national examinations. Working on this dilemma, this study sets out to design an assessment instrument for scientific argumentation competence (SAC) in Physics. The instrument is then used to explore Chinese high school students' current engagement in scientific argumentation.

Following a pragmatist orientation, both the students' overall performance on SAC assessment and their unique experiences of engaging in SA were explored to understand their SAC. Moreover, both the process and product of constructing the SAC assessment were examined through an iterative research design including four pilot studies and a main study. The initial theoretical framework for SAC was derived from a literature review on SA. This was then used to design initial SAC pencil and paper test items which were examined in the first pilot and then were modified to a complete SAC test to use in a second pilot. A third pilot was conducted to prepare the test for a large scale fourth pilot. The final version of the test was administered to 1413 students from seven schools in two regions of China, after which 12 students participated in the interview. Item response theory (IRT) was used to analyse test scores and thematic analysis was used to analyse interview data.

Findings highlight eleven factors to be useful in improving the assessment, such as scenario arrangement and provided information. By examining both the process and product of the assessment, the assessment shows acceptable validity, suggesting the rationality of assessing SA from the three components (i.e., Identification, Evaluation, and Production) and using the test scores. Reflecting on the inconsistencies between the IRT results and initial assumption leads to a possible three-level learning progression of SAC. This is aligned with and further expands previous learning progressions of SA. Most of the students are at level 1 of the learning progression. Students' assessment performance differs across schools and classes and has weak positive relationships with their school achievement test scores in Physics and Chinese. This study also found that despite the students' unfamiliarity with SA, knowing the definition of SA elements does not lead to better performances on the test. Interview data shows that students were positive about the idea of integrating SA into teaching and learning of Physics. However, they were pessimistic regarding the practical likelihood of this ever being implemented.

Discussion of the findings highlights the hybrid nature of SA and possible learning progression(s) of SAC, thus contributes theoretically to the possible ways of framing SA. This study advances previous research by providing a guideline for designing SA assessments and therefore, it contributes with theorising the assessment of SA. This study also contributes to knowledge on Chinese high school students' perceptions, experience, and performance on SA. Implications are drawn for the ways to demonstrate SA in school science/Physics curricula and to improve the students' SA engagement, highlighting that assessing SA in high stakes examinations, while only part of the effort, is critical for integrating SA into classrooms.

## Acknowledgements

I would like to thank the lovely people who have supported this doctoral research and helped me through the PhD journey. It is the care, kindness, companionships, and ideas offered by these people that makes this PhD study possible.

My first and deepest thanks go to my supervisors, Dr Bill Browne, and Dr Angeline Barrett. I could not have accomplished this tremendous task without your support. The PhD journey has not been easy, but because of you, it has never been bitter. Your critical feedback throughout this journey has not only mobilized ideas on writing this thesis, but also inspired me in terms of how to be a researcher and an educator. I am deeply grateful to both of you, you have always been there with your insightful suggestions and friendly encouragement whenever I needed support. I will forever be grateful for the trust you have been giving me, to me, you are the best supervisors in the world.

I would also like to thank the friends and colleagues I met in Bristol for your sustained encouragement and critical suggestions. Many thanks to Dini, Betzabé, Carolina, Artemio, and Nidia. You have provided helpful feedback on the thesis and have offered me valuable enlightenment along the way. Thank you for pushing me to broaden the experience during my PhD and for encouraging me to think over many other possibilities that have been so helpful for enhancing my PhD experience. I would also like to express my thanks to SoE colleagues who have given me mental support throughout this journey. SoE has been a community full of love, happy, support, and solidarity.

My sincere gratitude also goes to Prof George Leckie and Prof Guoxing Yu for having given me feedback that are critical and intellectually stimulating to organize my thesis. Thanks to Prof Alf Coles and Dr Philippa Howard for trusting me to join in your team to do research. Thanks to my examiners Prof Jo-Anne Baird and Prof Shelley McKeown Jones for giving me an amazing Viva experience. Thanks to Prof Michael Reiss for your kind and support during my visiting, which has been an extremely encouraging experience and I learned from you what a scholar and gentleman should be like.

I am also indebted to all the participants of this doctoral research and the teachers and friends who have offered help during the data collection. To all the lovely students for your trust in me, and for your interest and contribution to this study.

My heartfelt thanks also go to my dear friends in China. Although we are not working in the same field, the solidarity and critical enlightenment entailed in the friendship have supported me to deal with the PhD and my life positively. You have made me a better person and made me believe that.

Great thanks to UoB and China Scholarship Council for funding this four years' PhD research, without which I would not have the great learning experience at University of Bristol. Thanks to Bristol Collegiate Research Society for sponsoring me to the conferences and for UKRI for the COVID-impact funding. You have provided such big support so that I could have concentrated on my doctoral research.

Finally, my special thanks go to my beloved family. Great thanks to my mom Guohuan Zhao and my dad Changjian Zhang, your selfless love has nourished my heart to be powerful. Great thanks to my two little lovely sisters, Jingyue Zhang and Jingxuan Zhang, for calling me almost every day and sharing your tiny happy/boring things to make me feel happy or at least less bored. You have given me so much tolerance for my sometimes impatient and childish personality. You have not only been the pillar for these four years, but also for my whole life. I had been, am, and will love you unconditionally, as you have been doing, through my life.

## **Publication statement**

Some of the material included in Chapters 6 and 8 has been published in Zhang, J., & Browne, W. (2022). Exploring Chinese High School Students' Performance and Perceptions of Scientific Argumentation by Understanding it as a Three-components Progression of Competences. *Journal of Research in Science Teaching*.

Some of the material included in Chapter 5 will be published in Zhang, J., & Browne, W. An Approach to Generating Guidelines for Designing Scientific Argumentation Competence Assessments. *Contributions from science education research: Fostering Scientific Citizenship in an Uncertain World - Selected Papers from the ESERA 2021 Conference*.

### **Author's declaration**

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Jinglu Zhang

DATE: 10<sup>th</sup> July 2022

## Table of contents

<b>Abstract</b> .....	<b>i</b>
<b>Acknowledgements</b> .....	<b>ii</b>
<b>Publication statement</b> .....	<b>iii</b>
<b>Author’s declaration</b> .....	<b>iv</b>
<b>Table of contents</b> .....	<b>v</b>
<b>List of figures</b> .....	<b>11</b>
<b>List of tables</b> .....	<b>13</b>
<b>Acronyms</b> .....	<b>15</b>
<b>Chapter 1. Introduction</b> .....	<b>1</b>
1.1 Rationale for the research.....	2
1.1.1 Academic rationale .....	2
1.1.2 Local rationale .....	3
1.1.3 Personal rationale .....	4
1.2 Research aims and questions.....	4
1.3 Theoretical overview .....	5
1.4 Overview of methodology.....	6
1.5 Thesis overview.....	7
<b>Chapter 2. The Research Context</b> .....	<b>9</b>
Introduction .....	9
2.1 Curriculum reforms in China .....	9
2.2 The exam-oriented culture in China.....	12
2.3 The Gaokao examination .....	14
Chapter summary .....	16
<b>Chapter 3. Literature Review</b> .....	<b>18</b>
Introduction .....	18
3.1 Assessment, teaching, and learning.....	18
3.1.1 Key terms related to educational assessment .....	18
3.1.2 Impact of educational assessment.....	19
3.2 Conceptual understanding of SA and SAC .....	22
3.2.1 What is scientific argumentation? .....	22
3.2.2 SA as epistemic practice.....	24
3.2.3 Scientific argumentation competence.....	26
3.3 Assessments for scientific argumentation .....	29
3.3.1 The structure of SA.....	30

3.3.2	The content of SA.....	31
3.3.3	The epistemic aspects of SA.....	32
3.3.4	Exploring SA as a learning progression .....	35
3.3.5	Guidance for developing SA assessments .....	37
3.3.6	Assessment format for SA.....	39
3.4	Students' SA engagement .....	40
3.4.1	SA and content knowledge .....	41
3.4.2	SA and instruction .....	43
3.4.3	SA and cultural context .....	45
3.5	Assessment validation .....	47
3.5.1	Argument-based approach for validation .....	48
3.5.2	The macro- and micro-validation perspectives .....	50
3.5.3	Including test-takers' voices into validation.....	51
	Chapter summary .....	52
<b>Chapter 4.</b>	<b>Methodology .....</b>	<b>55</b>
	Introduction .....	55
4.1	Philosophical underpinning.....	55
4.2	Research design.....	58
4.3	Sampling.....	63
4.3.1	Overall consideration.....	63
4.3.2	Sampling for Pilot I to Pilot III.....	64
4.3.3	Sampling for Pilot IV .....	65
4.3.4	Sampling for the main study.....	67
4.4	Data collection.....	69
4.4.1	Data collection strategy .....	69
4.4.2	Data collection procedure.....	74
4.5	Data analysis .....	78
4.5.1	Think aloud data analysis .....	78
4.5.2	Semi-structured interview data analysis .....	81
4.5.3	Test data analysis.....	82
4.5.4	Descriptive statistics .....	85
4.5.5	Inferential statistics.....	85
4.6	Ethical considerations .....	86
4.6.1	Researcher access and informed consent.....	86
4.6.2	Anonymity.....	88
4.6.3	Participants' right and experience .....	88
	Chapter summary .....	90
<b>Chapter 5.</b>	<b>Developing a SAC Assessment.....</b>	<b>91</b>
	Introduction .....	91
5.1	The four building blocks for constructing measurement .....	91
5.1.1	Construct map.....	92



5.1.2	Items design .....	93
5.1.3	Outcome space.....	93
5.1.4	Measurement model .....	94
5.2	Assessment design I- An initial attempt.....	94
5.2.1	Construct map I .....	96
5.2.2	Test version I .....	98
5.2.3	Scoring rubrics I .....	101
5.3	Assessment design II- A rich exploration .....	102
5.3.1	Construct map II .....	102
5.3.2	Test version II.....	103
5.3.3	Scoring rubrics II.....	108
5.4	Test version III- Abridged and focused.....	108
5.4.1	Main findings from pilot II.....	109
5.4.2	Test modification .....	112
5.5	Test version IV- Finishing touches .....	114
5.5.1	Main findings from pilot IV .....	115
5.5.2	Test modification.....	116
	Chapter summary .....	117
<b>Chapter 6. Validating the SAC Assessment and Understanding the SAC Construct</b>		<b>119</b>
	Introduction .....	119
6.1	IUA of the assessment.....	119
6.2	Validity argument for the micro-validation .....	124
6.2.1	Claim 1 The instrument development procedure produces items that elicit SAC 124	
6.2.2	Claim 2 The test administration follows the prescribed procedure .....	137
6.2.3	Claim 3 The scoring process is consistent and accurate for all examinees .....	138
6.2.4	Summary.....	139
6.3	Validity argument for the macro-validation.....	139
6.3.1	Claim 4 The internal structure of the construct is represented accurately in the assessment .....	139
6.3.2	Claim 5 There is no negative impact on the participants by implementing the assessment .....	146
6.3.3	Summary.....	147
6.4	Further consideration about the underperforming test items .....	148
6.4.1	Items Erb_7.2 and Ee_7.3.....	148
6.4.2	Items Prb_4.3 and Prb_5.3 .....	150
6.4.3	Summary.....	151
6.5	Further consideration about the construct of SAC .....	151
6.5.1	Implications for a learning progression of SAC.....	152
6.5.2	Chinese high school students' performance on the SAC learning progression... 158	
	Chapter summary .....	160

<b>Chapter 7. Students' SAC Performance and Relevant Factors.....</b>	<b>162</b>
Introduction .....	162
7.1 Data subsets.....	162
7.2 Context and gender.....	163
7.2.1 Area .....	163
7.2.2 School.....	164
7.2.3 Class.....	165
7.2.4 Gender .....	166
7.3 Scaffold .....	168
7.4 Content knowledge.....	169
7.4.1 School 5 .....	169
7.4.2 School 6 .....	170
7.4.3 School 3 .....	171
Chapter summary .....	173
<b>Chapter 8. Understanding Students' SAC from Their Perspective .....</b>	<b>175</b>
Introduction .....	175
8.1 Participants .....	175
8.2 Students' perceptions about SA .....	177
8.2.1 Existing awareness of SA transferred from previous experience.....	177
8.2.2 Positive attitude on SA and the assessment.....	180
8.3 Students benefit from taking the SAC assessment.....	183
8.3.1 Pedagogical function of the assessment .....	184
8.3.2 Introspections through the assessment .....	185
8.4 Challenges of engaging in SA.....	188
8.4.1 Lack of opportunities to engage in SA .....	188
8.4.2 Difficulties of engaging in SA.....	191
Chapter summary .....	194
<b>Chapter 9. Discussion.....</b>	<b>196</b>
Introduction .....	196
9.1 The nature of scientific argumentation.....	196
9.1.1 Understanding SA from a competence perspective.....	196
9.1.2 Understanding SAC as a learning progression .....	204
9.2 Equipping students with SAC by assessing it.....	208
9.2.1 Chinese high school students' SAC.....	209
9.2.2 What to expect from assessing SAC?.....	212
9.3 Developing assessments for scientific argumentation .....	215
Chapter summary .....	221
<b>Chapter 10. Conclusion.....</b>	<b>224</b>

Introduction .....	224
10.1 Answering research questions .....	224
10.1.1 RQ1. How can a SAC assessment be designed for high school Physics students in China?.....	224
10.1.2 RQ2. To what extent is the developed SAC assessment valid and reliable for assessing SAC?.....	225
10.1.3 RQ3. What does the developed SAC assessment provide in terms of extended understanding of SA and of Chinese high school students' SAC?.....	225
10.1.4 RQ4. How does the SAC of Chinese high school students as measured by the SAC assessment differ between different student groups? .....	226
10.1.5 RQ5. What are Chinese high school students' perceptions of SA and the challenges they face in SA engagement? .....	226
10.2 Implications .....	227
10.2.1 Implications for policy .....	227
10.2.2 Implications for science teaching.....	229
10.3 Contributions.....	231
10.3.1 Contribution to knowledge .....	232
10.3.2 Contribution to methodology .....	234
10.4 Limitations and future research.....	235
<b>COVID-19 statements.....</b>	<b>239</b>
<b>Reference .....</b>	<b>240</b>
<b>Appendix 1 Comparison of Curriculum 2003 and 2017.....</b>	<b>258</b>
<b>Appendix 2 Ferrara and Lai's (2015) validation framework .....</b>	<b>262</b>
<b>Appendix 3 Semi-structured follow-up interview outline .....</b>	<b>265</b>
<b>Appendix 4 Nvivo coding screenshot of think aloud data .....</b>	<b>266</b>
<b>Appendix 5 Theme construction drafts.....</b>	<b>267</b>
<b>Appendix 6 Nvivo coding screenshot of follow-up interview data .....</b>	<b>270</b>
<b>Appendix 7 Codes and themes of thematic analysis .....</b>	<b>271</b>
<b>Appendix 8 Participant information sheet for students (think aloud interview).....</b>	<b>282</b>
<b>Appendix 9 Participant information sheet for students (test and follow up interview) .....</b>	<b>284</b>
<b>Appendix 10 Participant information sheet for teachers .....</b>	<b>286</b>
<b>Appendix 11 Students consent form for participation in research .....</b>	<b>288</b>
<b>Appendix 12 Teachers consent form for participation in research.....</b>	<b>289</b>
<b>Appendix 13 SoE research ethics form .....</b>	<b>290</b>
<b>Appendix 14 Test version I-teacher.....</b>	<b>296</b>
<b>Appendix 15 Scoring rubrics I.....</b>	<b>300</b>
<b>Appendix 16 Test version I-students .....</b>	<b>302</b>
<b>Appendix 17 Test version II-teachers .....</b>	<b>306</b>

<b>Appendix 18 Test version II-students .....</b>	<b>315</b>
<b>Appendix 19 Test version III .....</b>	<b>323</b>
<b>Appendix 20 Test specification .....</b>	<b>331</b>
<b>Appendix 21 Test version IV.....</b>	<b>333</b>
<b>Appendix 22 Scoring rubrics III.....</b>	<b>339</b>

## List of figures

Figure 1.1 Research overview.....	7
Figure 3.1 The three components of SAC .....	29
Figure 3.2 Toulmin’s Argument Pattern (Toulmin, 1958) .....	30
Figure 4.1 How Pragmatism understands the world.....	56
Figure 4.2 Mixed methods research design .....	61
Figure 4.3 Locations for data collection in China.....	64
Figure 4.4 Sampling approach for the fourth pilot .....	66
Figure 4.5 Sampling approach for the main study .....	68
Figure 4.6 Test papers posted to and collected back from schools in Jilin.....	77
Figure 5.1 A generic construct map in construct “X” (Wilson, 2004) .....	92
Figure 5.2 Instrument development procedure .....	95
Figure 5.3 P-SA-Explanation task example (Test version I) .....	99
Figure 5.4 I-SA task example (Test version I).....	100
Figure 5.5 E-SA-use of evidence task example (Test version I) .....	101
Figure 5.6 Task example (Test version II).....	107
Figure 5.7 Scaffold and I-SA task 1 (Test version III) .....	113
Figure 5.8 E-SA task 1 (Test version III) .....	114
Figure 5.9 P-SA task 1 (Test version III).....	114
Figure 5.10 E-SA task 2 (Test version IV) .....	117
Figure 5.11 Strategies employed to improve the SAC assessment.....	118
Figure 6.1 Micro and Macro validation for the SAC assessment .....	120
Figure 6.2 The claims network of the IUA for the SAC assessment.....	122
Figure 6.3 Scree plot in fourth pilot.....	131
Figure 6.4 ICC/CCC of items in the fourth pilot .....	135
Figure 6.5 Test information curve (fourth pilot).....	136
Figure 6.6 Wright map (main study).....	142
Figure 6.7 Test information curve (main study) .....	143
Figure 6.8 CCC plots for task 4 .....	144
Figure 6.9 CCC plots for task 5 .....	144

Figure 6.10 CCC plots for task 6 .....	145
Figure 6.11 CCC plots for task 7 .....	145
Figure 6.12 Empirical plot for Erb_7.2.....	149
Figure 6.13 Empirical plot for Ee_7.3 .....	149
Figure 6.14 CCC for Prb_4.3 (before and after revision) .....	150
Figure 6.15 CCC for Prb_5.3 (before and after revision) .....	151
Figure 6.16 Wright map showing the progression levels of SAC .....	157
Figure 6.17 Distribution of the students' SAC .....	160
Figure 7.1 Scatterplot of school Physics scores in school 5 .....	170
Figure 7.2 Distribution of SACT scores and school Physics scores in school 5 .....	170
Figure 7.3 Scatterplot of SACT and school Physics scores in school 6 .....	171
Figure 7.4 Distribution of school Physics scores and SACT scores in school 6 .....	171
Figure 7.5 Scatterplot of SACT and school Physics and Chinese scores in school 3.....	172
Figure 7.6 Distribution of school Physics scores and school Chinese scores in school 3 .....	172
Figure 7.7 Distribution of SACT scores in school 3.....	172
Figure 9.1 Understanding of SAC revealed by this study .....	203
Figure 9.2 SA assessment guideline .....	221

## List of tables

Table 2.1 Key competences in 2017 Physics curriculum .....	10
Table 2.2 Achievement progression in the Physics curriculum for high school education .....	11
Table 2.3 Grade goals in the Science curriculum for compulsory education .....	11
Table 3.1 Analytical framework of Erduran et al. (2004).....	31
Table 3.2 Deng and Wang's (2017) framework .....	32
Table 3.3 Epistemic levels for argumentation analysis (Kelly & Takao, 2002).....	33
Table 3.4 Lee et al.'s (2014) assessment framework for SA .....	34
Table 3.5 Osborne et al.'s (2016) learning progression.....	36
Table 3.6 IUA proposed by Kane (2006, 2009).....	49
Table 3.7 Interpretation argument proposed by Shaw and Crisp (2012).....	49
Table 4.1 Sample of the first three pilots .....	65
Table 4.2 Sample of the fourth pilot .....	67
Table 4.3 Sample of the main study.....	69
Table 5.1 Construct map I.....	98
Table 5.2 Scoring rubric example (Test version I- task 3) .....	101
Table 5.3 Construct map II .....	103
Table 5.4 Scoring rubric example for Pr items .....	108
Table 5.5 Scoring rubric example for Prb items .....	108
Table 6.1 Claims in the IUA of the SAC assessment .....	121
Table 6.2 Item difficulty estimates (fourth pilot).....	128
Table 6.3 Item-total correlation .....	130
Table 6.4 Item parameters in the fourth pilot .....	134
Table 6.5 Model data fit and items estimates (main study) .....	141
Table 6.6 Items estimates (after modifying Prb_4.3 and Prb_5.3) .....	153
Table 6.7 Learning progression of SAC .....	155
Table 6.8 Measures for each SAC level.....	158
Table 7.1 Area difference in SAC performance .....	164
Table 7.2 School difference in SAC performance .....	165
Table 7.3 Class type and SAC performance .....	166

Table 7.4 Gender difference in SAC performance .....	167
Table 7.5 Scaffold and SAC performance .....	169
Table 8.1 Interview sampling .....	177
Table 9.1 Comparison of learning progressions .....	205



## Acronyms

CCC	Category characteristic curve
CTT	Classical test theory
E-SA	Evaluation of scientific argumentation
Ee	Evaluation of evidence
Er	Evaluation of reason
Erb	Evaluation of rebuttal
ICC	Item characteristic curve
Ie	Identification of evidence
Ir	Identification of reason
IRT	Item response theory
I-SA	Identification of scientific argumentation
IUA	Interpretation/Use argument
LR	Likelihood ratio test statistic
PCM	Partial credit model
P-SA	Production of scientific argumentation
Pe	Production of evidence
Pr	Production of reason
Prb	Production of rebuttal
RQ	Research question
SA	Scientific argumentation
SAC	Scientific argumentation competence
SACT	Scientific argumentation competence test
SSI	Socio-scientific issues

## **Chapter 1. Introduction**

Over the last few decades, the focus of science education has been shifting from “what we know to how we know and why we believe” (Duschl, 2008, p.269). Hence, argumentation as a core element of both scientific practice and scientific thinking (Khine, 2011), is being introduced into school science education around the world (NGSS, 2013; ACARA, 2016; DfE, 2014; Ministry of Education, P. R. China, 2017). Scientific argumentation (SA) encompasses the construction, evaluation, and reconstruction of scientific knowledge (Osborne et al., 2016; Ford, 2012). Scientific argumentation engages students in activities such as proposing and supporting claims, using evidence, and evaluating each other’s ideas and reconciling their differences (González-Howard & McNeill, 2020). The skills involved in SA of making sense of information are increasingly valued in today’s world where information is oversupplied and knowledge can be obtained easily (Gilbert, 2005). However, in many countries including China, argumentation seldom happens in science classrooms (Berland & Hammer, 2012; Szu & Osborne, 2012; Deng & Wang, 2017). There is still a scarcity of knowledge about the nature of the competences involved in SA for researchers and especially teachers and curriculum developers (Khine, 2011). Therefore, to change the current state of school science education, it is necessary and valuable to further explore the nature of SA competence (SAC) and its development.

Argumentation is usually investigated in the form of dialogic discourse in which people are interacting with each other (Berland & Reiser, 2011), or in the form of individual discourse where the conversation happens internally (Sampson & Clark, 2008). Assessing and analysing either form of argumentation has been an important way for researchers to understand the nature of SA. Moreover, teaching and learning is heavily influenced by assessments (Cheng & DeLuca, 2011), most especially high-stakes tests like the Gaokao in China. Gaokao is the national college entrance exam in China, which is a curriculum-based high-stakes test that tests high school students’ mastery of the subjects taught in school (Bai et al., 2014). It is the most critical examination for most Chinese youth and parents (Muthanna & Sang, 2015). Over 10 million high school students took the test in 2021 (Ministry of Education, P. R. China, 2021). SA is likely to continue to be omitted from teaching and learning in high school science in China as long as SAC is not assessed in examinations, despite its explicit inclusion in the new Curriculum document of high school Physics.

Hence, assessing SA helps understand the nature of SA, and designing SAC assessments that

can be administered under examination conditions is a useful starting point for expanding the possibilities for science teaching and learning. Additionally, exploring how Chinese high school students engage in SA helps further understand SA and its practice in China.

## **1.1 Rationale for the research**

### **1.1.1 Academic rationale**

In the last 20 years, researchers across the world have shown an increasing interest in investigating argumentation in science education (Erduran et al., 2004; Berland & McNeil, 2010; Osborne et al., 2016; Chen & Qiao, 2020). Previous studies have highlighted the important role of SA that can play in helping students enhance their understanding of scientific concepts, develop their views on the nature of science, and improve their scientific literacy (Khishfe, 2014; Cavagnetto, 2010; Zohar & Nemet, 2002). However, there are still emerging studies that have shown, in the context of SA, that students' ability to use evidence, reasoning and especially rebuttal of arguments is poor (Hogan & Maglienti, 2001; Clark & Sampson, 2005; Lee et al., 2020; Cavagnetto et al., 2010; Deng & Wang, 2017). In order to engage students in SA, the assessment of SA has been an important research effort in addition to its teaching (Lee et al., 2014; Sampson & Clark, 2008). Currently, researchers have developed a variety of approaches to theorising argumentation but "no clear and homogeneous definition exists for argumentative competence and its constituent skills" (Rapanta et al., 2013, p. 483). This has brought challenges for its assessment in terms of how differing theoretical frameworks challenge how researchers communicate findings and how to assess various aspects of SA in a valid and reliable fashion (Henderson et al., 2018).

Due to this, research in SA assessment has been shifting from assessing / analyzing either the process of SA discourses or the argument product students generate to exploring SA as learning progressions and assessing students' development of competences on these progressions from less skilled to more skilled (Osborne et al., 2016; Berland & McNeill, 2010; Lee et al., 2014; Ng Yee Ping, 2019). **Thus, going beyond nuanced and diverse analysis of SA to investigate the assessment of SA in a more comparable, comprehensive, and easy-for-practice way can advance the understanding and assessment of SA.**

Assessments that use quantitative methods and can be administered at large scale remain scarce (Osborne et al., 2016), and most of the assessments developed are more suitable for research purposes as they are complicated and time-consuming for teachers to conduct (Dawson &

Carson, 2017). Correspondingly, there have been few studies discussing how to design instruments to assess SA and the challenges that may be faced in the process (Ng Yee Ping, 2019). Most studies either talk about the results of assessing SA, or the instrument as a product. Consequently, there is a conspicuous lack of research investigating possible guidelines that could be used in SA assessment so as to make it a sustained joint effort between researchers in the area of SA assessment. Guidelines for assessments do not only “formalize the content and cognitive specifications of items but also automate their development” (Shute et al., 2016, p. 51). There thus may be reciprocal causal relationships between the lack of SA assessment guidelines, the very many ways of analyzing or assessing SA, and the absence of SA assessment and teaching in science classrooms. **Therefore, exploring SA assessment in a systematic and sustainable way and making SA assessment transparent and well-documented can facilitate the construction/adaption and application of SA assessments.**

Overall, the academic rationale for this study is to contribute to the aforementioned research gaps, to further explore how SA can be understood and assessed.

### **1.1.2 Local rationale**

In line with the emphasis on key competences in the latest curriculum reform initiated in 2015, for the first time, scientific argumentation was explicitly included in the new curriculum for high school Physics and the new curriculum for elementary and middle school science (Ministry of Education, P. R. China, 2017, 2022). Specifically, SA was proposed as an element for one of the key competences (i.e., scientific thinking) in the science/Physics curricula. Details of Chinese context and the curriculum reform will be introduced in Chapter 2. **So, exploring SA and its assessment meets the needs of China’s curriculum reform, and can provide further implications for policy.**

Despite the large number of studies around the world, there is very limited research on SA in China (Deng, 2015; Zhang, 2018; Zheng, Zhang, & Zhang, 2019). Most studies in China focus on theory and attempt to call for emphasis on this area by reviewing studies conducted in other countries (Deng & Wang, 2014; Dong, 2018; Han, Hu, & Wang, 2014; Ren & Li, 2012; Song & Wang, 2018; Wu & Liu, 2017). Moreover, scholars have discussed and questioned how Chinese students who are influenced by the ‘harmony’ culture might engage in SA (Xie et al., 2015; Spencer-Rodgers et al., 2010; Osborne et al., 2016). **So, it is of great value to investigate how Chinese students engage in scientific argumentation to add to the**

**literature and thus update knowledge.**

### **1.1.3 Personal rationale**

My personal interest in exploring this topic can be traced back to my own high school experience in China. We saw little relevance of learning science to our real life and lifelong development. At that time, science especially Physics was the most difficult subject for most students. Teachers usually told us that Physics is a subject that is relevant to our daily life, but when we came across science phenomenon outside of school, it was hard to connect them with the knowledge we had learnt at school. We had a superficial understanding of the nature of science. Learning Physics had been more about remembering and applying formulas and providing answers to problems using the steps the teacher taught us. Gradually, I realized there was an absence of critique and argument in science education in China which prevented students' deep thinking concerning what is science. So, I conducted my master's research on a topic related to SA. Due to the limited time and experience, there were still many questions that remained unsolved and so I decided to continue my study to further explore this area. Further several of my friends, who were former classmates at university, became school Physics teachers. They all aspire to help their students to improve their ability to reason and argue, and so obtain a deeper understanding of Physics phenomena and concepts. However, they know little about SA and how to teach and assess SA.

In addition, I would also like to improve my own SAC through the process of conducting this study. In this study, I plan to explore SA in the context of Physics because my own educational background is in Physics. Nonetheless, choosing Physics as the context is not only because of my own background, but also for the policy rationale as mentioned in the previous section.

## **1.2 Research aims and questions**

This study aims to design and evaluate a pencil and paper test for assessing scientific argumentation competence (SAC) and to understand Chinese high school students' engagement in SA. By doing this, this study sheds light on the nature of SA, the way to assess SA, and the feasibility for Chinese high school students to acquire SAC. This study explores scientific argumentation in the context of the subject of high school Physics, a subject where SA is a specified learning outcome in the curriculum. To achieve this aim, the objectives of this study are to:

- 1) review the international literature on scientific argumentation, including its conceptualisation, its assessment and students' engagement in SA;
- 2) explore the construction, modification, and validation of a pencil and paper scientific argumentation competence (SAC) assessment that can be administered at large scale;
- 3) investigate the possible factors that may be associated with students' performance on the SAC assessment;
- 4) explore the students' experience of sitting the SAC assessment and their experience of learning science in school; and
- 5) draw out implications for the theory, policy and practice relating to SAC and its assessment and teaching.

Specific questions that guide the empirical research are:

***RQ1.** How can a SAC assessment be designed for high school Physics students in China?*

***RQ2.** To what extent is the developed SAC assessment valid and reliable for assessing SAC?*

***RQ3.** What does the developed SAC assessment provide in terms of extended understanding of SA and of Chinese high school students' SAC?*

***RQ4.** How does the SAC of Chinese high school students as measured by the SAC assessment differ between different student groups?*

***RQ5.** What are Chinese high school students' perceptions of SA and the challenges they face in SA engagement?*

### **1.3 Theoretical overview**

This research is guided by theories in the field of argumentation and educational assessment. This study mainly draws on Toulmin's (1958) argument pattern (TAP) and Erduran et al.'s (2004) adaption of TAP to understand the structure of SA. Additionally, this study draws on previous research concerned with developing argumentation competence (Kuhn et al., 2013; Rapanta et al., 2013). Previous studies in terms of understanding SA as an epistemic practice, concerning with argument evaluation, and assessing SA were reviewed to theorize SAC (e.g., Osborne et al., 2016; Sandoval, 2003). These thus lead to a framing of SAC that breaks it down into three constituent components, namely **identifying SA**, **evaluating SA**, and **producing SA**.

With respect to educational assessment, this study is informed by theories in both assessment development (Wilson, 2004) and assessment validation (e.g., Newton, 2017; Kane, 2013) to make sense of both the product and process of assessment. Different understanding of SA itself leads to various ways of assessing it. Thus, to assess a construct such as SA, it is important to

clarify the construct. Wilson's (2004) approach of developing an assessment guides this study for its emphasis on understanding and evaluating the construct to be assessed and for its specific guidance for practice. In addition, this study combines an argument-based validation approach (conduct validation by formulating validity arguments based on empirical evidence obtained during developing an assessment) (Kane, 2013) and a Micro-Macro validation approach (conduct validation by evaluating an assessment product and the process of developing it) (Newton, 2017) to illuminate the evaluation of both the product and process of the SAC assessment.

#### **1.4 Overview of methodology**

To understand Chinese high school students' SAC, this study focuses on both general information obtained from large scale assessment and students' unique experience of sitting the assessment and learning school science. Therefore, a pragmatist philosophical position is embodied in this study (Morgan, 2014a). Considering the focus of both the process and product of developing a SAC assessment and the interest of exploring SA by assessing it and probing into students' experience, a mixed methods methodology is adopted in this study (Johnson et al., 2007). To achieve the aim of constructing a SAC assessment, an iterative research design is employed, which is consistent with Dewey's (1949) elaboration of the acting, thinking, and reflection entailed in experience.

Specifically, an initial version of the SAC assessment is designed based on the initial conceptualization of SAC, four pilot studies are conducted to modify and improve the assessment with the results of each pilot contributing to the implementation of the next. Qualitative data manifesting from the interaction between teachers and targeted students with the assessment, and quantitative large-scale test data (in the fourth pilot) are analysed to inform *the appropriate ways of designing a SA assessment and the validation of the assessment process*. The final version of the SAC assessment generated based on the pilot studies is administered at large scale to inform *the validation of the assessment product, the further reflection of the SAC construct, and Chinese high school students' overall performance*. The results derived from the above stages address RQ1, RQ2, and RQ3, which mainly point to SAC measurement.

The students' performance obtained from the validated SAC assessment is then used to analyse the impact on SA of SA related variables, such as school, class, gender, assessment scaffold

(i.e., definition of SA) etc. RQ 4 is addressed at this stage to *uncover the factors that may influence the students' SAC performance thus to further understand SA*. The follow-up interviews conducted after the two large-scale studies are analysed to answer RQ 5 with the aim to *understand SA and its implementation in classrooms from students' perspectives*.

As shown in Figure 1.1, the above stages together address the overarching research aim of *designing and evaluating a pencil and paper test for assessing scientific argumentation competence (SAC) and understanding Chinese high school students' engagement in SA* and provide insights in terms of the nature of SA, how to help Chinese high school students acquire SAC, and the assessment of SA.

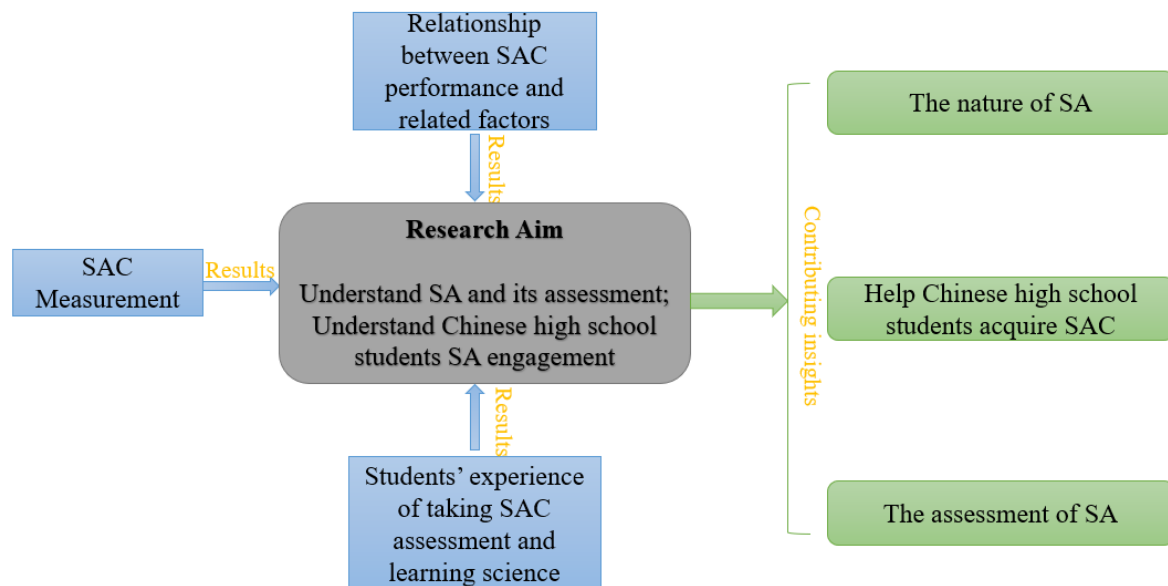


Figure 1.1 Research overview

## 1.5 Thesis overview

This thesis consists of 10 chapters. The current introduction chapter has provided an overview of the study as a whole, outlined the rationale of conducting the research, and introduced the research aim and questions. Chapter 2 sets the research in its contemporary context, introducing the curriculum reforms in China to highlight that the inclusion of SA in the curricula is not accidental, but an inevitable result of the curriculum reform history. Chapter 2 also illustrates the exam-oriented culture and the Gaokao examination in China to elaborate why exploring the assessment of SA is important for school education in China.

Chapter 3 reviews the international literature in two fields, namely SA and educational



assessment. For SA, literature about its conceptual understanding and assessment, students' engagement in SA and related influencing factors are reviewed. By doing so, the chapter sheds light on how SA is approached in this study and understanding Chinese high school students' perception of SA and their assessment performance. For educational assessment, literature about the role of assessment in education and assessment validation are reviewed to situate assessment in the overarching background of education and to inform the implementation of assessment in this research.

Chapter 4 sets out the philosophical position and research design of this study. The chapter first illustrates why and how Pragmatism is used as the philosophical orientation of this study. Then, the iterative mixed-methods research design is described, including detailed explanation of methods of data collection and analysis and discussion of ethical considerations.

Chapters 5 to 8 present research findings. Chapter 5 shows how the SAC assessment was developed and modified through iterative piloting. Research question 1 (RQ 1) will be addressed in the chapter. Chapter 6 moves onto validation for the SAC assessment. The Chinese high school students' performance on the assessment is discussed as well. The chapter addresses RQ 2 (construct validity of the assessment) and RQ 3 (expanded understanding about SA and the students' SAC). Chapter 7 addresses RQ 4. It presents results from the large-scale administration of the assessment tool, investigating relationships with variables such as gender, by location, student performance in Physics and Chinese and scaffold within the assessment tool. Chapter 8 presents findings from thematic analysis of student interviews regarding their experience of participating in the SAC assessment, to address RQ 5.

Chapter 9 discusses the findings in Chapters 5 to 8 to shed light on the overarching aim of this research, namely, *to design and evaluate a pencil and paper test for assessing scientific argumentation competence (SAC) and to understand Chinese high school students' engagement in SA*. Evidence related to the nature of SA, how to help Chinese high school students acquire SAC, and the construction of SA assessment are compared and discussed. The thesis will be concluded in Chapter 10 by summarizing the answers for each research questions, providing implications for policy and science teaching, summarizing academic contributions, and considering limitations and possibilities for future research.

## **Chapter 2. The Research Context**

### **Introduction**

This chapter aims to elaborate the rationale for the research by setting the research in its contemporary context. Section 2.1 will introduce the policy background that prompted me to undertake this research, namely the recent curriculum reforms that introduced an emphasis on SA into the specified curricula for science. However, there currently seems to be a gap between the presence of SA in the curriculum and its implementation in school education, possibly due to the absence of SA in examinations. Thus, section 2.2 will demonstrate the exam-oriented culture that has influenced Chinese education system for thousands of years and was an indirect reason for the recent educational reforms. Section 2.3 will provide an overview to Gaokao which is an important part of the current Chinese education system and is a significant examination that further enhances the exam-oriented culture.

### **2.1 Curriculum reforms in China**

Since the founding of the new China in 1949, the national curriculum for school education in China has experienced several rounds of reforms. The most significant ones happened in 1978, 2001, and 2015 respectively. The curriculum issued in 1978 prioritised learning of basic knowledge and basic skills. Curriculum reform in 2001 paid more attention to students' comprehensive development (i.e., knowledge & skills; process & methods; emotions, attitudes & values), aiming to transform from exam-oriented education to quality-oriented education (Yan, 2015). The 2015 curriculum reform responded to an international agenda promoting 21<sup>st</sup> century skills as a priority for education (Xin, 2016). So currently, the focus of the school curriculum in China is shifting towards promoting students' key competences rather than only knowledge to lay the foundation for students' comprehensive and lifelong development of literacy.

It was as part of this shift towards defining learning outcomes in terms of competences rather than merely knowledge acquisition, that SA was first explicitly included in the high school Physics curriculum, released in 2017 (see Table 2.1). However, teacher-centred practices remained dominant in Chinese classrooms after the curriculum reforms in 2001 and 2015 (OECD, 2010, 2020). Hence, Chinese students were less proficient in interpreting data and evidence scientifically than in reproducing content knowledge and explaining phenomena scientifically (OECD, 2020).

Table 2.1 Key competences in 2017 Physics curriculum

Key competences for subject Physics	Elements
Physics conceptual understanding	Substance; movement and interactions; energy
Scientific thinking	Model construction; scientific reasoning; <b>scientific argumentation</b> ; question and innovation
Scientific inquiry	Question; evidence; explanation; communication
Scientific attitude and responsibility	Nature of science; attitude of science; social responsibility

The national curriculum is the basis for textbook writing, teaching, assessment, and examinations (Ministry of Education, P. R. China, 2017, 2022). Although detailed prescription of teaching content in the syllabus issued in 1952 left less flexibility for teaching (Li, 2009; Cui, 2001), the lack of specific instructions in recent decades' curriculum could be one of the reasons that curriculum reforms did not achieve their objective. Another reason may be the influence of examinations which will be introduced in the next section. By comparing the Physics curriculum issued in 2003 (based on the curriculum reform in 2001), and that issued in 2017 (based on the curriculum reform in 2015), it seems that what has changed is not the goals themselves but how the goals are specified and can be achieved. Namely, the previous Physics curriculum also stated "learning the knowledge and skill needed for lifelong development, promoting scientific thinking, and learning how to do scientific inquiry to solve problems, etc." as its goals. In terms of the involved competences, as shown in Appendix 1, most of the key competences that were proposed in the 2017 curriculum existed in the 2003 curriculum as well. However, the key competences are listed in the 2017 curriculum explicitly rather than only implied by the goals, and in addition, a five-levels achievement progression (see Table 2.2 for a part of the progression related to SA) is proposed in the 2017 curriculum to show how these competences develop from less skilled to proficient. That is, a general trend in the curriculum reform in China, as has been pointed out in the new curriculum, is to provide more specific instructions for teaching, learning and assessment in the curriculum to support realizing the overarching goals.

Table 2.2 Achievement progression in the Physics curriculum for high school education

Levels (From low to high)	Scientific thinking
Level 1	...; Can distinguish claim and evidence;
Level 2	...; Can express one's claim using simple and direct evidence;
Level 3	...; Can express one's claim by using evidence appropriately;
Level 4	...; Can justify Physics conclusion by using evidence appropriately;
Level 5	...; Can consider the reliability of evidence;

Nonetheless, the new curriculum does not elaborate on the nature of SA and provide enough information to guide its teaching, learning and assessment. Likewise, SA is manifested in the achievement progression in a rather simple way (Table 2.2). Both in the progression and in the overview of key competences, SA is described together with other elements in Scientific thinking such as scientific reasoning. The curriculum thus does not provide systematic and comprehensive elaboration for each sub-competence on their own.

Recently in April 2022, the Ministry of Education released new curricula for compulsory education (i.e., primary school and secondary school) and included the same four key competences in the science curriculum as that in the Physics curriculum issued in 2017 (but replacing 'Physics conceptual understanding' with 'Science conceptual understanding'). Instead of proposing an achievement progression for each competence, the Science curriculum proposes goals for each grade level to achieve. Table 2.3 below shows the SA related part as demonstrated in the curriculum.

Table 2.3 Grade goals in the Science curriculum for compulsory education

Grades	Scientific thinking
Grade 1-2	...; under teacher's instruction, students can preliminarily distinguish claim and facts, and have an awareness of providing evidence;
Grade 3-4	...; under teacher's guidance, students can build the connection between claim and facts, propose hypothesis, and provide supportive evidence;
Grade 5-6	...; students can propose hypothesis, propose their claim when communicating with others, and build connections between evidence and claim;
Grade 7-9	...; students can test hypothesis and reach conclusions based on evidence and logic, interpret the plausibility of their claim, and engage in rebuttal based on evidence.

Interestingly, the newly released Science curriculum for compulsory education seems to have provided a more elaborated description of SA, focusing on building *connections* between claim and facts and explicitly mentioned *rebuttal* (see section 3.3.1 for the exact definition of 'rebuttal'). However, the curricula for primary education, secondary education, and high school

education do not seem to show progressively increasing demanding on students' SA. In other words, the SA included in the high school curriculum does not build on that in the compulsory education curriculum. And still, the curricula for compulsory education does not provide specific separate explanations in terms of the nature even the definition of SA and the guidance for its teaching and assessment. This may lead to teachers' insufficient sense of the conceptualizations and differences in these competences and thus impede their endorsement of the curricula, which could further impede the implementation of the curricula in classrooms (Yu et al., 2018).

Overall, it is a positive signal for the new curricula in China that they include SA explicitly, indicating the rationality and necessity of paying attention to SA. However, SA is not manifested in the curricula in a systematic and elaborative way. Therefore, further research and practice in terms of the theoretical understanding and educational practice of SA are needed.

## **2.2 The exam-oriented culture in China**

Another possible reason why recent curriculum reforms have fallen short of their goals is that the exam-oriented culture has far-reaching influence on the Chinese education system. Examinations have a long history in China as a narrow competitive gateway to opportunity for further education or status, which can be traced to the 7<sup>th</sup> century in the Sui dynasty, namely China's civil exam system or Keju (Yu & Suen, 2005). Keju was proposed to select competent officials to serve the emperor and create unified value within the nation to control the society and culture (Chen et al., 2020; Li & Wang, 2022). Keju was open to all males in ancient China including those from commoner backgrounds, and it was relatively free of corruption while extremely competitive (Chen et al., 2020). As the only way for commoners to change social mobility (Gan, 2002), the rewards of being successful in Keju were extraordinary, including considerable income, prestige, and privileges (Li & Wang, 2022). The extremely high returns of passing Keju may be one of the reasons that promoted a culture of valuing education among the Chinese (Chen et al., 2020). As illustrated in a famous poem, named "Urge to Study" written by Emperor Zhenzong (真宗) (968-1022 AD) of the Song dynasty (Yu & Suen, 2005):

富家不用买良田，书中自有千钟粟。  
安房不用架高堂，书中自有黄金屋。  
出门莫恨无随人，书中车马多如簇。

娶妻莫恨无良媒，书中自有颜如玉。

男儿欲遂平生志，六经勤向窗前读。

*“To be wealthy you need not purchase fertile fields, Thousands of tons of corn are to be found in the books.*

*To build a house you need not set up high beams, Golden mansions are to be found in the books.*

*To travel you need not worry about not having servants and attendants, Large entourages of horses and carriages are to be found in the books.*

*To find a wife you need not worry about not having good matchmakers, Maidens as beautiful as jade are to be found in the books.*

*When a man wishes to fulfil the ambition of his life, He only needs to diligently study the six classics by the window.” (p. 18)*

Despite the aim of Keju being to elect competent officials, it was focused on examining the mastery of Confucianism. Although Confucius had always advocated the importance of education and practiced it throughout his whole life, the way people have practiced his educational doctrines has deviated from his original intent. It is recorded in the Analects that “仕而优则学，学而优则仕” (“An official should devote his leisure to learning if he has discharged all his duties; If it is easy for a person to have a high moral character and do things in an appropriate way, he should become an official to help more people and serve the society”). However, with the implementation of Keju and under the context that being an official means having power and wealth, people have begun to only care about the latter part of the proverb and mistaking it as “As long as a person succeeds in the Keju, he can become an official and thus gain prestige” (Luo, 2005). Education thus has been endowed with more utilitarian functions and become associated with family honour in practice for a long history, which brought ‘exam-driven education fever’ and accompanying negative impacts (Yu & Suen, 2005). These impacts include promoting the polarization of society, narrowing what is valued by society and school, and harming students’ mental health etc. (Yu & Suen, 2005).

A similar form of examination in contemporary China is the National College Entrance Exam (NCEE) or Gaokao, which is widely believed to be a critical step for social mobility (Yu & Suen, 2005). Upper secondary education (in which students are aged around 16 to 18) is almost entirely directed towards preparing students for Gaokao, and the goal of primary and lower secondary school education is to enter key high schools with outstanding performance in Gaokao (Wang et al., 2020). Although now Gaokao is not the only way to achieve upward

mobility, it is a primary way for millions of high school students to acquire education in good universities, live a better future, gain honour for their family (Wang et al., 2020). Gaokao has been described as a “single wooden pole bridge” that everyone wishes to pass successfully (Liu & Helwig, 2020, p. 3). Except for Gaokao, other high-stakes examinations are an important part of education experience for Chinese K-12 students (OECD, 2020). The exam-oriented culture thus has been passed down to today, in which teachers and students devalue what would not be tested (Yu & Suen, 2005), and teachers could not change how they teach even after the issue of the curriculum in 2001 because they “can’t change much unless the exams change” (Yan, 2015, p. 5). In addition, private tutoring had been expanding rapidly in China to improve students’ scores in Gaokao, until the release of the ‘Double reduction’ policy by the government in July 2021. ‘Double reduction’ refers to ‘Further Reducing the Homework Burden and off-campus training Burden of Students in Compulsory Education’, aiming to address not only educational burden on students but also problems caused by the exam orientation such as the social class solidification, parents’ anxiety on education, high spending on education, students’ physical and mental burden (Xue & Li, 2022).

Overall, the exam-oriented culture has existed in China for a long time and has had profound impacts on the school education in China and Chinese society. The negative impacts caused by such orientation have been widely recognized, but it would not be a one-time job to diminish such impact. Even so, changing what and how to examine would be an inevitable part of this tremendous endeavour.

### **2.3 The Gaokao examination**

Gaokao is administered once a year in June, and the overall test score on it is nearly the only determinant of whether and what kind of university a student goes to (Liu & Helwig, 2020). The Gaokao has six subjects in total, with three compulsory subjects (i.e., Chinese, mathematics, English) and three other subjects in Chemistry, Physics, Biology, History, Geography, and Politics. Before the reform of Gaokao that has been gradually implemented in different provinces from 2014, high school students needed to choose between science-stream subjects (i.e., Chemistry, Physics, Biology) or arts-stream subjects (i.e., History, Geography, and Politics). Now, students can choose any combination of three subjects based on their own willingness.

Gaokao has been found to some extent to be perpetuating social inequalities although some

argue that it has benefits such as motivating study (Yu et al., 2018) and enhancing equity and credibility of college enrolment (Liu & Wu, 2006). The disparity in education between rural and urban areas of China has been found to be manifested in school facilities, teacher quality, rate of dropouts from lower secondary school and high school, scores obtained in the Gaokao, the mastery of content knowledge, cognitive ability, family support, and opportunities to attend extra learning activities (Wang & Peng, 2012; Ye, 2011; Zhang et al., 2015; Liu & Helwig, 2020; Liu, 2013). However, students from both urban and rural areas of China expressed similar views in discussing the aim of education and their current learning experiences driven by the Gaokao (Liu & Helwig, 2020). Specifically, students see the aim of education as related to stimulating interest in learning, deeper understanding of knowledge, appreciation of life, and human flourishing (Liu & Helwig, 2020; Li, 2006). However, students tend to complain about the repetitive nature of preparing for Gaokao and connect high school, Gaokao, and even college study with extrinsic values and motivations such as a means of getting a well-paying job (Liu & Helwig, 2020; Muthanna & Sang, 2016). Despite this, students in general do not support the abolition of Gaokao and think that the existence of Gaokao makes them study hard (Liu & Helwig, 2020).

There are other reasons that lead to critiques of the Gaokao than simply its high-stakes nature. Students tend to describe Gaokao as a ‘memory-based test’ that does not emphasize creative and critical thinking (Muthanna & Sang, 2016; Liu & Helwig, 2020). The design and implementation of Gaokao are all arranged and monitored by the Ministry of Education of the People’s Republic of China, including national unified propositions implemented by the Ministry of Education and provincial propositions implemented by educational organizations in each province. Professionals who are mostly teachers in universities are selected by the educational organization in each province to design the Gaokao test items. The Gaokao test items are different in different provinces but are all aligned with the national curriculum and Gaokao examination outline (Ministry of Education, P. R. China, 2006). Along with the awareness of the critiques to Gaokao and the recent reforms in curriculum and Gaokao, the Ministry of Education issued the ‘Evaluation system of the Chinese National College Entrance Exam’ (hereinafter referred to as ‘The evaluation system’) in 2020. The evaluation system proposes the concepts of ‘One core, Four layers, and Four wings’. ‘One core’ is the core function of Gaokao, namely ‘cultivating morality, serving talents, guiding teaching’, answering the question of ‘why the examination’. ‘Four layers’ are the content of Gaokao, that is, ‘core values, subject literacy, key competence, and necessary knowledge’, answering the question of



‘what to examine’. ‘Four wings’ are the requirements for Gaokao, that is, ‘basic, comprehensive, applied, and innovative’, and answer the question of ‘how to examine’. By comparing the Physics Gaokao test in different provinces in 2021 and previous years, it has been found that test items have started to include more real-life scenarios and scenarios related to real science research (Dai & Xu, 2021; Zhou et al., 2022; Zhang & Zhang, 2022). As in the previous years, test items in Gaokao require students to have solid understanding on content knowledge and high reasoning skills (Xia et al., 2022; Li et al., 2022), while still requiring students to provide a final correct answer and do not include argumentation explicitly. Nevertheless, open-ended items that require students to construct their explanations have started to appear in the test (Meng et al., 2022).

Since the issue of the new Physics curriculum in 2017, there has been a conspicuous lack of empirical evidence in terms of whether SA has been integrated into the Physics classroom. Zhang and Chen (2018) found that high school Physics teachers in China in general perceive SA as an important practice, but they had a weak understanding about SA and its teaching, and they seldom practice SA in their classrooms. Given the lack of research exploring SA in China and the prevailing teacher-centred classroom in China (Yu et al., 2018; OECD, 2020), it seems reasonable to state that SA is generally absent from school Physics teaching and learning (Sun, 2019; Zhang & Chen, 2018). Combining with the previous review of studies, this may be due to the absence of SA in the Gaokao test items. However, it seems that now the design of Gaokao examination items is trying to assess students’ key competences and thus guide its teaching and learning. Therefore, it meets the trend of the recent educational reform in China to explore the assessment of SA. Furthermore, given that it is under-researched and under-practiced in the teaching and assessment in China compared to other competences such as scientific reasoning, studies on argumentation are required.

## **Chapter summary**

This chapter has laid out the background of this research. Based on the curriculum reforms, section 2.1 analysed that the aim of education has been shifting from focusing on knowledge to valuing the lifelong development of competence and literacy. As a result, SA has gained the attention of educational policy makers in China. However, the discussion and practice of SA in China is still in an early stage and the curricula for high school and compulsory education do not specify SA in a systematic way. So, research on SA is needed in China. Section 2.2 elaborated on how a long-lasting exam-oriented culture has been influencing the Chinese

education system and even Chinese society. Section 2.3 further justified the necessity of changing examinations by revealing the impact of Gaokao examination on teaching and learning and the absence of SA in Gaokao test items. The next chapter will further justify the importance of assessment by drawing evidence from studies internationally. Overall, it is of particular value to explore argumentation in the context of school education or school Physics education in China, especially to explore its assessment. So, the next chapter will review literature related to argumentation and SA in the context of school education.

## **Chapter 3. Literature Review**

### **Introduction**

This chapter aims to review literature related to SA and educational assessment to serve the research aim of exploring a SAC assessment and understanding Chinese high school students' SA engagement. To do this, this chapter first discusses why conducting assessment research is important for improving SA teaching and learning, then talks about the conceptual meaning of SA and how it has manifested in assessments. Then, with a focus of helping students acquiring SAC, how SA manifests and varies among students is reviewed. Lastly, this chapter discusses how an assessment of SAC can be validated.

In more detail, section 3.1 will first discuss the relationship between assessment, learning, and teaching. Section 3.2 will review the various theoretical understandings about SA and point to the importance of the three components (Identifying SA, Evaluating SA, Producing SA) in SA engagement and the rationale of framing SA as a competence. Section 3.3 will critically review the advantages and limitations of previous analytical/assessment frameworks to inform the assessment design of this study. These frameworks are organized based on their focus on the structure, content, epistemic aspects of SA and SA as learning progressions. Section 3.4 will review international research evidence on students' engagement in SA and how it relates to content knowledge, instruction, and cultural context. Section 3.5 will talk about the frameworks/methods that are usually used for assessment validation to inform the appropriate approach for this study.

### **3.1 Assessment, teaching, and learning**

This section clarifies the meaning of assessment and explores the role it plays in the overarching context of education to show how this study understands assessment and positions itself in the field of educational research.

#### **3.1.1 Key terms related to educational assessment**

In the field of education, assessment is the technique of collecting information relative to some known objective (e.g., knowledge, ability, attitude) (Kizlik, 2012; Scheerens et al., 2003). In contrast to testing and measurement, assessment has a broader meaning that entails more ways to assess students, in which both qualitative and quantitative information can be obtained,

interpreted, and used. Measurement is usually characterised as “the process of quantifying the observations (or descriptions) about a quality or attribute of a thing or person” (Thorndike & Hagen, 1986, p. 5). Testing is often used in the context of large-scale testing or high stakes testing that assesses subject matter knowledge or skills of a person, and test is often taken as one type of the instruments used for realizing assessment or measurement (Mohan, 2016). Literature on assessment usually talks about its purpose/function and that on measurement usually discusses its techniques and procedures (Newton, 2007; Scheerens et al., 2003). Overall, tests are assessments and have typically been used to mean educational measurement, but not all assessments need to be tests and not all measurements need to use tests (Mohan, 2016; Kizlik, 2012). Given this study aims to develop a pencil and paper test to assess Chinese high school students’ SAC, these terms are mostly used interchangeably. Generally, educational assessment serves the function of:

- 1) regulating the quality of educational outcomes and provisions,
- 2) accountability, and
- 3) stimulating improvement in education (Scheerens et al., 2003).

In contrast Newton (2007) proposes more categories for the use of educational judgement and argues that the primary purposes of an assessment should be made explicit when there are several functions for that assessment. Whatever the function, the purpose of assessment is always directed towards promoting quality of the targets at each level of educational systems, therefore assessment is an important part in educational system. For this study, the ultimate intention of doing this research is to help Chinese high school students to acquire the competence of engaging in SA. Creating an instrument to assess students’ SAC is an entry point for achieving this goal. Specifically, the purpose of the assessment in this study is to further refine the construct of SAC within high school Physics in a way that renders it measurable and to understand the students’ readiness to engage with SA. By so doing, this study tries to lay the foundations for advancing SA learning, teaching, and assessment in school education.

### **3.1.2 Impact of educational assessment**

Conducting educational assessment can have positive or negative, intended or unintended consequences for learning. The word ‘consequence’ refers more to the results of the use and misuse of assessment results and indicates a broader range of influence. ‘Washback’/

‘backwash’ and ‘impact’ refer to the influence of testing on teaching and learning, in which washback or backwash was first used in the field of applied linguistics and language testing (Cheng & Curtis, 2004). Thus, this study uses ‘impact’ to focus on the influence of assessment on learning and teaching. The remainder of this section will review existing studies that investigate the impact of educational assessment.

Educational assessment is usually considered to be what should and could drive teaching and learning, namely ‘measurement-driven instruction’ (Popham, 1987; Cheng & Curtis, 2004). Assessment has been found to enhance students’ performance and the retention of learned information (Roediger et al., 2011), and different assessment forms can have different impacts on students’ learning. For instance, McConnell et al. (2014) found that context-rich multiple choice question assessment and short-answer question assessment are significantly better at enhancing students’ learning than context-free multiple choice question assessment and studying alone without assessment. Giuliodori et al. (2008) found that collaborative group testing enhanced students’ performance compared to individual testing and students provided very positive feedback on the format of group testing. Appropriate forms of assessment can also support student-centred learning and teaching. By reviewing literature, Burner (2014) found that portfolio assessment allows students to do revision and reflection and increases students’ autonomy for learning. It has also been widely recognized that assessment can provide feedback for both students and teachers in terms of the strengths and weaknesses of teaching and learning thus to inform instruction for teaching and monitor students’ learning progress (Gallo et al., 2006). As mentioned in section 2.3, Chinese students found Gaokao made them study harder, thus large-scale testing was found to motivate both teachers and students to work harder and effectively (Abu-Alhija, 2007). In addition, assessment serves the role of conveying signals regarding what is desirable and what matters in education and in life (Abu-Alhija, 2007; Emler et al., 2019).

However, what is assessed and what is taught does not always reflect the needs of learners and the intentions of the curriculum (Qi, 2007; Yan, 2015). Negative impacts, especially those caused by high-stakes assessments, have been widely reported both in China and in other countries. Previous studies reported that high-stakes assessments impact teaching to be more exam-oriented thus leading to transmissive pedagogies (Amano & Poole, 2005; Polese et al. 2014). In China, Qi (2004) investigated the impact of the National Matriculation English Test (NMET), which is a high-stakes test for university entry, on the teaching in secondary school.

She found that the test shifted the goal of teaching to raising scores and focused teaching on test content, a focus that was narrower than the curriculum intended. Exam-oriented teaching often leads to the prevalence of teacher-centred practices and even memorization focused pedagogic methods (Yan, 2015). Similar findings are reported in studies conducted in other countries, where the main teaching activity is drilling the students with past papers as preparation for high-stakes tests rather than responding to students' learning needs (West, 2010; Polesel et al., 2014).

A limited number of studies paying attention to test-takers' views and experiences on assessments have emerged recently (Cheng & Deluca, 2011). Some of these studies aim at linking test-takers' experience to assessment validation, which will be discussed in section 3.5.3. Others reveal the impact of assessment, especially high-stakes assessment, on test-takers and their learning. High-stakes assessments narrow what students learn, impeding students' development, especially when the assessments focus on knowledge re-call. Test-takers in Cheng and Deluca's (2011) study talked about their experiences of taking large-scale English language tests in Asia. They expressed the importance of "tips and test-taking strategies" for obtaining a better score and showed their concern about the neglect of important competences that were not assessed by the test at the same time as declaring "test results" are "everything" (p. 114).

Widespread teaching and learning for the test shapes students' perception of and practice in educational assessments and learning in an undesired way. Such as leading to students lacking ability to apply knowledge in practice, low level of educational engagement, and focusing on test results over learning processes (Amano & Poole, 2005; Polese et al. 2014). Research has also demonstrated a mismatch between learners' learning strategies and the competences the assessments were intended to measure. Qi (2007) investigated high school students' perceptions of writing compared with those of the test constructors for the national matriculation English test in China. She found that whilst test-creator's intention was to encourage writing for communicative purpose, students aim at obtaining high scores rather than developing writing ability. Such findings have resonance in Andrews et al.'s (2002) study conducted in Hong Kong, which presents the impact of introducing an Oral examination to high-stakes tests on secondary school students' learning. They found that students tend to engage in superficial learning by familiarizing themselves with the exam format, rote-learning of exam-specific strategies and formulaic phrases, indicating memorisation rather than

“meaningful internalisation” (p. 220). Similarly, Cheng et al. (2011) conducted a study on how secondary school students in Hong Kong perceive a school-based English assessment which aimed at improving students’ Oral English proficiency. They found that students were not aware of what function the assessment was intended to have and took it as another exam like the ones they had taken in the past.

The above review indicates that both the intention and technique of assessment matter in educational research and practice since they do not only influence assessment itself but also teaching and learning. Therefore, it is of value to explore the assessment of SA and students’ experience of taking SA assessment and science learning to understand how the current assessments impact students’ learning and thereby impact on their SA engagement and how the students are influenced by engaging in SA assessment. These can further inform the possibility to improve Chinese high school students’ SA engagement during school education, and to reduce the negative impact, if any, of current assessments for the benefit of students. To design an assessment instrument however, it is first important to have a clear idea of what SA is. This is addressed in the next section.

### **3.2 Conceptual understanding of SA and SAC**

This study perceives SAC as the competences needed for successfully engaging in SA. Thus, this section explores different perspectives of understanding argumentation, particularly in science education, to show how SA is understood and framed in this research and why. By discussing argumentation as a *product* and *process*, as an *individual activity* or *social activity*, with the aim of *persuasion* or *collaboration*, and as *epistemic practice*, this section shows why and how SA is approached from a competence perspective.

#### **3.2.1 What is scientific argumentation?**

Giving a definition of SA is not straightforward and so instead of going directly to the meaning of SA, a brief review of argumentation is needed. This section starts by clarifying the distinction between *argument* and *argumentation*. Although there are various different usages of terms in the field of argumentation, “argument” has often referred to a *product in which one or more assertions are supported by evidence and justifications of evidence toward an issue or topic* (Angell, 1964; Halpern & Vardi, 1989; Khine, 2011; Means & Voss, 1996; Schwarz et al., 2003; Toulmin, 1958; Zohar & Nemet, 2002). In contrast “argumentation” usually refers to a *process during which people have interactions with each other and pay attention to each*

*other's argument* (Rapanta et al., 2013; Kuhn & Udell, 2003). Despite differences in the definitions for argumentation, the core content remains similar, taking argumentation as “a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of one or more propositions to justify this standpoint” (Van Eemeren et al., 2002, p. xii). In general, as many previous educational research studies do, this study uses “argument” as *the artifacts that a student or a student group creates when articulating and justifying their claims*, whereas using “argumentation” as *the process by which such artifacts are produced* (Ryu, 2011; Sampson & Clark, 2008; Kuhn & Udell, 2003; Jiménez-Aleixandre & Erduran, 2007).

Within the existing literature, argumentation is usually allocated with either **individual meaning** or **social meaning**. The **individual meaning** of argumentation refers to the internal process, for example, when people articulate and justify their claim in their article or speech without interacting with other people directly (Billig et al., 1988; Yang et al., 2015). From this perspective, scholars use “monological argumentation” to indicate situations where a single person constructs their arguments through the process of implicit dialogues which takes place in one’s mind (Goldman, 1999; van Eemeren et al., 1984; Yang et al., 2015). Some scholars use “rhetorical argumentation” to represent the individual meaning of argumentation where the audience does not interact with the arguer directly, emphasizing the intra-personal process (Blair, 2012; Vygotsky, 1987). In terms of the **social meaning** of argumentation, Jimenez-Aleixandre and Erduran (2008), agreeing with Billig (1987), describe it as “a dispute or debate between people opposing each other with contrasting sides to an issue” (p. 12). This understanding focuses on the goal of **persuasion** for argumentation that aims to undermine opposing arguments and persuade others. While some scholars highlight that argumentation is “the interaction with other people to reach an agreement by putting forward a series of propositions to justify the claim” (Ryu, 2011; Van Eemeren et al., 2013, p. 5). This concern with agreement indicates the goal of **collaboration** for argumentation that focuses on negotiation and reaching a consensus (Leitão, 2000; Ryu, 2011; Noroozi et al., 2013).

The different perspectives of understanding argumentation show that not only the activity of argumentation as a whole has its purpose (i.e., collaboration and persuasion), but the statements that form an argumentation have their functions as well. Thus, to be engaged in argumentation, people need to know the function that each statement has and be able to generate statements that serve desired functions. Therefore, this study takes **identifying** the function of each



statement in an argumentation and **producing** statements that have a certain function as two important components of engaging in argumentation.

This study is concerned with SA. Literature on SA seldom discusses its definition but take it as the argumentation that takes place in the context of science or when engaged in scientific topics or issues (Erduran et al., 2004; Driver et al., 2000). However, the disciplinary context can have implications for how an argumentation can be framed. Jiménez-Aleixandre and Erduran (2008) define argumentation in scientific topics as “the connection between claims and data through justifications or the evaluation of knowledge claims in light of evidence, either empirical or theoretical” (p. 13). Many researchers have found it important for students to think and talk like scientists, while SA is an important process when scientists are working on scientific issues or problems (Duschl, 2008). Thus, the evidence and reasoning used in the process or presented in the argument should be reasonable and scientific, evidence could be obtained through the investigation of the natural world or the implementation of scientific experiments. Lee et al. (2014) also proposed that people should be able to identify the uncertainty in scientific problems or claims and produce alternative claims.

Taking all the ideas into consideration, this study defines SA as *a product generated through an internal or social process of using scientific evidence to defend one’s scientific claims reasonably, meanwhile using scientific evidence to evaluate the strengths and weaknesses of others’ claims.*

### **3.2.2 SA as epistemic practice**

To engage in SA, people need to understand what counts as argumentation and what counts as good argumentation in science so as to understand what and how to argue to generate knowledge (Ryu & Sandoval, 2012; Duschl, 2008; Chen et al., 2019). SA is an epistemic practice in and of itself and engaging in such practice can lead to an understanding toward the epistemological base of scientific practice (Sandoval & Millwood, 2007). Epistemology in science refers to “the study of the growth of knowledge, the nature of evidence, the criteria for theory choice and the structure of disciplinary knowledge” (Kelly, 2008, p. 99), or “beliefs about the nature of science and scientific knowledge” (Sandoval, 2003, p. 8). Epistemic practice is *the specific ways to propose, justify, evaluate, and legitimize knowledge within a certain discipline by specific community members* (Kelly, 2008, p. 99). Argumentation is a central practice of science and engaging in SA is a process of making a series of “what counts”

epistemic judgements, such as what counts as evidence and what counts as coherent argument (Duschl, 2008). Moreover, engaging in SA entails negotiating conflicting ideas based on how the science community constructs knowledge, such as identifying and evaluating *uncertainty* triggered by “conflicting, incomplete, and diverse knowledge claims and evidence” (Chen, & Qiao, 2020, p. 2) or attending to *critique* by challenging and evaluating some aspects of other’s arguments (González-Howard & McNeill, 2020). Therefore, SA is an epistemic practice closely related to epistemological understanding of science.

On the one hand, previous studies have revealed that the enhanced epistemic understanding of SA can lead to better SA performance. For example, Nussbaum et al. (2008) conducted a study aimed at undergraduate students suggesting that students who received information of the nature of sound scientific arguments tended to incorporate more scientific criteria into their discussion and developed better arguments. Other studies have also shown that students’ argumentation quality can be improved through instructions that exposed students to the criteria used to construct and evaluate an argument (Duschl & Osborne, 2002; Nussbaum et al., 2005).

On the other hand, the development of epistemic understanding of SA and epistemological beliefs about knowledge and knowing are reciprocal and are shaped mutually (Kuhn et al., 2000; Lee, Liang, & Tsai, 2016). Realists and the absolutists view knowledge from external sources and are certain, multiplists have an awareness of uncertain and take claims as subjective opinions that are freely chosen and equally right. However, evaluativists view that one position can have more merit if it is better supported by evidence and argument (Kuhn et al., 2000). Previous studies found that students with a sophisticated epistemological belief (e.g., evaluativists) are more willing and capable to negotiate uncertainty, evaluate different arguments, and appreciate multiple solutions when learning scientific knowledge (Reznitskaya & Gregory, 2013; Reznitskaya et al., 2009). Additionally, students with an evaluativist epistemology belief tend to generate higher quality arguments (Mason & Scirica, 2006). In a reciprocal way, engaging in argumentation has been found to lead to more sophisticated epistemological beliefs (Iordanou, 2010; Ryu & Sandoval, 2012).

All in all, SA is a social practice in which members of the science community constitute a set of actions based on common purposes with shared tools and meanings (Gee & Green, 1998), it is also an epistemic practice that conforms to consistent ways of justifying and evaluating knowledge. Gaining epistemic understanding of SA is an intrinsic part of the SA practice that

supports better SA engagement and the development of epistemological understanding of science, thereby supporting students' science learning. Therefore, following the two components mentioned in section 3.2.1 (i.e., **identifying** and **producing** scientific arguments), **evaluating** scientific arguments, by which students' epistemic understanding of SA is externalized, is also an important component of SA engagement.

### 3.2.3 Scientific argumentation competence

This study argues that explicating the competences needed for SA engagement could facilitate designing SA assessments that demonstrate SA comprehensively, are easy to implement in classrooms and provide instructional information for teachers. A competence consists of a set of dispositions that are necessary for coping with certain situations or problems, and observable performance that can result from its elicitation (Koeppen et al., 2008; Weinert, 2001; Blömeke et al., 2015). Studies that explore high order thinking skills in science education and explicitly talk about competence often take competence as a context-specific construct that requires different dispositions depending on the situation and can be trained and required. For example, Rapanta et al. (2013) considers argumentative competence as “the ways in which different types of skills related to argumentation are manifested in a person's performance” (p. 488). Wang and Song (2021) explicitly discuss interdisciplinary competence and take competence as “the internal structure of competence in terms of basic abilities” (p. 694), and Reith and Nehring (2020), in a study of scientific reasoning, view competence as “dispositions that are acquired and needed to successfully cope with certain situations or tasks” (Koeppen et al., 2008, p. 62). There are few studies in the field of SA talking about SA competence explicitly despite SA being a competence-based discourse that has been understood in various ways (as discussed in section 3.2.1 and 3.2.2). SA competence is thus an ill-defined concept that is underexplored (Rapanta et al., 2013).

As shown in section 3.2.1, the literature talks about the **theoretical meaning** of SA from various perspectives. However, there should be ways to synthesize these perspectives since they are talking about the same thing, namely argumentation. To have a closer look at these perspectives, the *product* of argument is generated through the *process* of argumentation either *internally* or *socially*; and the *process* of argumentation should be able to generate an argument as a *product* regardless of whether it's truly generated or not. No matter what *goal* is assigned to argumentation, its practice needs to go through a *process* and have the potential to generate a *product*. Although a certain goal of argumentation might bring more benefits to participants

if it has the potential to elicit more about reflection, reconciliation, and reconstruction (Felton et al., 2009; Felton et al., 2015; Garcia-Mila et al., 2013; Evagorou & Osborne, 2013; Kuhn, 2015). Thus, all these perspectives together form the SA entity.

Based on different perspectives of SA, **empirical studies** of SA use various frameworks of analysis and with different foci (Rapanta et al., 2013). Many analytical frameworks of SA have enriched our understanding about argumentation in science education, whilst highlighting the challenge of how to make studies based on different understandings comparable and how to choose between these understandings for explicit guidance for science teaching (Henderson et al., 2018; Quinlan, 2020). Most of the empirical studies investigate either the quality of student-generated arguments or the dynamic processes of argumentation activities using various frameworks (which will be illustrated in section 3.3) that capture either the characteristics of an argument or the behaviour presented in an argumentation. However, as previously mentioned, different perspectives of SA should be able to be synthesised, this study further argues that the characteristics presented from an argument product and the behaviour demonstrated in an argumentation process should all be supported by competences related to SA engagement. In other words, the acts and the abilities needed for engaging in argumentation do not change as often as these frameworks, although some forms of argumentation put higher and-or explicit demands on certain abilities and other forms have higher and-or explicit requirements for other abilities (Henderson et al., 2018).

Literature that talks about SA from a competence perspective usually provides broader insights of how to analyze it. Kuhn et al. (2013) argued that developing argument competence is multifaceted and does not only contain the production and evaluation of argument encompassing metacognitive, epistemological, and social dimensions, but also involve dispositions, values, and norms. Rapanta et al. (2013) proposes a three-tier conceptualization of argumentative competence: meta-cognitive, meta-strategic and epistemological. But neither of these studies provides empirical evidence to justify their conceptualization. Therefore, approaching SA from a competence perspective could be a way to realize a comprehensive understanding and assessment of it, therefore, to generate explicit and specific instructions for teaching.

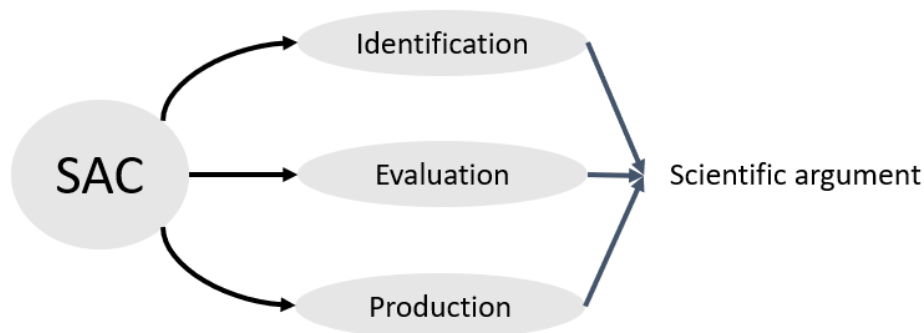
So, how could this study frame SAC? The evaluation of an argument plays an important role in science learning and argumentation activities, although some existing studies analysing SA emphasize the ability of formulating an argument and justifying it (Clark & Sampson, 2005;

Yang et al., 2015). From the perspective of how scientific knowledge is constructed, it is widely accepted that it is socially constructed through “critique, replication and evaluation” (Dawson & Carson, 2017, p. 4). Moreover, the ability to evaluate new ideas in a critical, reflective and rational manner is valued by today’s society (Osborne, 2014). As for SA itself, as mentioned previously, SA is an epistemic practice in which knowledge needs to be evaluated. Kuhn (1993) referred to argumentation skill as the development of logical explanations and recognition of opposing claims, weights of evidence, and determination of the merit of each claim based on evidence. Lastly, evaluation serves a pivotal function in the process of argumentation (both individually and socially). Rebuttal and counterargument are two important elements of argumentation that indicate higher quality of argumentation (Erduran et al., 2004; Osborne et al., 2016; Sampson & Clark, 2008). However, if taking a further step into these two elements of argumentation, the premise of proposing reasonable counterargument and rebuttal is knowing the strengths and weaknesses of their own and others’ argument and why these strengths and weaknesses exist, this indicates the process of evaluation. The significance of evaluation can also be found in Walton’s (1989) claim that there are two goals for skilled argumentation, one is to “secure commitments from the opponent that can be used to support one’s own argument”, and the other is to “undermine the opponent’s position by identifying and challenging weaknesses in his or her argument” (Walton, 1989, cited from Kuhn and Udell, 2007, p. 91). Overall, evaluating arguments allows students to distinguish and/or make judgements about the quality of arguments based on the understanding of what should be good arguments (Britt et al., 2014).

Recent empirical studies have started to pay attention to argument evaluation, although most studies are concerned with teachers. Martín-Gómez and Erduran (2018) found that teachers’ ability to evaluate arguments was weak. Using the same instrument, Zhao et al. (2021) found that Chinese preservice science teachers’ abilities to evaluate and construct SA had significant and moderate correlation between them. Lytzerinou and Iordanou (2020) asserted that science teachers’ ability to construct arguments predicted their ability to evaluate arguments. Glassner et al. (2005) found that students can successfully evaluate the plausibility of statements in supporting explanation and argumentation while showing difficulty in generating argumentation. Thus, despite recognizing that the abilities to evaluate and construct arguments are related, the existing studies seem unable to determine which is a prerequisite. As for the empirical studies on argument identification, Von der Mühlen et al. (2016) and Münchow et al. (2019) found that college students’ evaluation of arguments has a significant positive

correlation with their identification of functional argument elements. They thus conjecture that argument evaluation skills might be improved by fostering argument construction skills. It was also conjectured that competences involved in evaluating argument and systematically identifying argument elements might be part of a common construct (Von der Mühlen et al., 2016; Britt et al., 2014). Similarly, Larson et al. (2009) found that teaching students the structure of argument improved their argument evaluation.

In conclusion, as an ill-defined area (Rapanta et al., 2013), the discussion and empirical evidence in existing studies seem to preliminarily support the assumption that the competences of **identifying** a scientific argument (I-SA), **evaluating** a scientific argument (E-SA), and **producing** a scientific argument (P-SA) are related and essential components for engaging in SA. These three components therefore conceptualize the understanding of SAC in this study (see Figure 3.1).



*Figure 3.1 The three components of SAC*

### 3.3 Assessments for scientific argumentation

The different perspectives on understanding SA lead to various ways to analyse/assess it, and the different analysis/assessment frameworks further broaden the understanding of SA. Considering this study is focused on assessing argumentation using pencil and paper assessment, this section will not review the assessment frameworks including social interactions. Although some assessment frameworks were designed to analyse the argumentations that occur in social interactions, they usually also encompass other dimensions. So, the categorization below depends on whether the framework itself can shed light on the discussion about a pencil and paper assessment regardless of the context in which it was designed. In addition, the categorisation below captures the main features of each framework, but overlaps exist between them. Literature shedding light on how to design a pencil and paper

SA assessment will be discussed at the end of the section.

### 3.3.1 The structure of SA

An influential model in the field of argumentation is Toulmin's argumentation pattern (TAP) which is concerned with the procedure of eliciting different elements of an argument (i.e., 'claim', 'data', 'warrant', 'backing', 'qualifier', and 'rebuttal') that have different logical functions (Toulmin, 1958; Nielsen, 2013). Toulmin (1958) suggested that arguments can be classified into six elements by their function (see Figure 3.2): a **claim** is "*an assertion put forward publicly for general acceptance*". **Data** are "*the facts explicitly appealed to as a foundation for the claim*". **Warrants** are "*the specific facts relied on to support a given claim*". **Backings** are "*generalizations making explicit the body of experience relied on to establish the trustworthiness of the ways of arguing applied in any particular case*". **Qualifiers** are "*phrases that show what kind of degree of reliance is to be placed on the conclusions, given the arguments available to support them*" and **rebuttals** are "*the extraordinary or exceptional circumstances that might undermine the force of the supporting arguments*" (cited from Erduran et al., 2004, p. 918). From Toulmin's perspective, the quality of an argument is based on the presence or absence of these elements, where arguments with more of these elements would be stronger. Although there are several drawbacks of Toulmin's argument model that have been recognized by researchers in the area, for example, many studies have suggested that it is hard to distinguish between Toulmin's elements especially data, warrants and backings (Erduran et al., 2004), it still plays a significant role in the understanding and conceptualization of an argument.

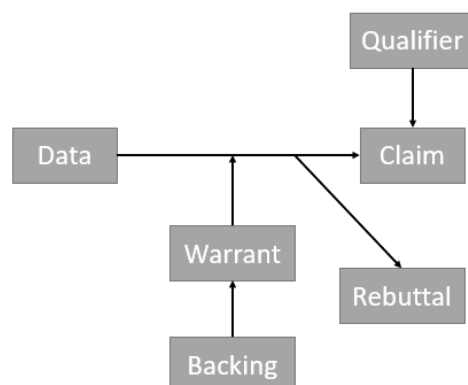


Figure 3.2 Toulmin's Argument Pattern (Toulmin, 1958)

Researchers in science education have either applied TAP or modified it in studies dealing with

written arguments or argumentation discourse for its advantage in understanding the structural dimensions of argumentation (Osborne et al., 2016; Deng & Wang, 2017; Nielsen, 2013; Sampson & Clark, 2006, 2008; González-Howard & McNeill, 2020). Among these studies, a significant framework is the one put forward by Erduran et al. (2004). This framework is a modification of TAP aiming at fixing its weaknesses and was applied in the science classroom. Erduran et al. (2004) merged data, warrants and backing into one category labelled reason and their framework takes arguments with rebuttals as high-quality argumentation. Erduran et al. (2004) illustrated 5 levels to assess the quality of argumentation as shown in Table 3.1.

*Table 3.1 Analytical framework of Erduran et al. (2004)*

<b>Level 1</b>	Argumentation consists of arguments that are a simple claim versus a counterclaim or a claim versus a claim.
<b>Level 2</b>	Argumentation has arguments consisting of a claim versus a claim with either data, warrants, or backings but do not contain any rebuttals.
<b>Level 3</b>	Argumentation has arguments with a series of claims or counterclaims with either data, warrants, or backings with the occasional weak rebuttal.
<b>Level 4</b>	Argumentation shows arguments with a claim with a clearly identifiable rebuttal. Such an argument may have several claims and counterclaims.
<b>Level 5</b>	Argumentation displays an extended argument with more than one rebuttal.

Erduran et al.'s (2004) study was not focused on the assessment of argumentation but rather on the analysis of students' argumentation discourse in science classrooms. However, their study provided a new perspective on categorizing SA into different levels by identifying key aspects of argumentation and emphasizing the importance of rebuttal, which can be used as the basis of a framework for assessing the quality of an argumentation. The above studies provide insights on **understanding the structure of SA and the complexity of different SA elements**, although they provide little information about students' ability to distinguish argumentation quality.

### **3.3.2 The content of SA**

The content of SA here emphasizes the character of the statements in an argument, instead of focusing on the existence and quantity of the functional statements as studies that are concerned with the structure of SA do. Zohar and Nemet's (2002) study analysed both 9<sup>th</sup> grade Israeli students' writing and discussion related to human genetics and is significant for its emphasis on justification and discipline knowledge. The goal of their study was to explore the effects of integrating explicit teaching of general reasoning patterns into the teaching of biological knowledge on both the resulting biological knowledge and argumentation skills. For the analysis of written arguments, this study also put Toulmin's data, warrants and backings into a



single category and considered arguments with multiple justifications as high-quality arguments. They analysed students' ability to formulate arguments, alternative arguments, and rebuttals by counting the number and complexity of justifications (valid justification, simple justification, complicated justification). As for biological knowledge, their focus was on whether and to what extent would students consider biological knowledge when constructing their arguments. The criteria for knowledge in argumentation included four levels: No biological knowledge is considered, Incorrect consideration of biological knowledge, Consideration of non-specific biological knowledge and Correct consideration of specific biological knowledge. However, as the authors admitted in their study, the analysis is rather simple and only offers a partial criterion for argumentation quality.

Deng and Wang (2017) proposed a framework based on Toulmin's argument model to assess Chinese high school students' written SA in the context of Chemistry. They assigned 2 to 4 levels of performance for the elements of SA (i.e., claim, evidence, warrant and rebuttal). Their framework analyzed the quality of each element of an argument by considering whether they are scientific, sufficient, and articulated. However, they did not consider the differences between these elements rather applying the same criteria for all elements.

*Table 3.2 Deng and Wang's (2017) framework*

<b>Structure components</b>	<b>Level 1</b>	<b>Level 2</b>	<b>Level 3</b>	<b>Level 4</b>
<b>Claim</b>	No claim	Scientifically correct		
<b>Evidence</b>	No evidence	Scientifically correct	Scientifically correct and sufficient	Using detailed, precise, and unambiguous language based on level 3
<b>Warrant</b>	No warrant	Scientifically correct	Scientifically correct and sufficient	Using detailed, precise, and unambiguous language based on level 3
<b>Rebuttal</b>	No rebuttal	Refuting one claim effectively	Refuting all claims effectively	Using detailed, precise, and unambiguous language based on level 3

By reviewing these studies, we can see that the analysis of SA that focuses on its content are usually concerned with either the discipline knowledge or the argumentation as a practice per se. The above studies therefore shed light on **assessing an argument by considering both content knowledge and the nature of argumentation.**

### **3.3.3 The epistemic aspects of SA**

Studies exploring the epistemic aspects of SA have focused on either the scientific knowledge

entailed in an argument or argumentation as a scientific practice per se. As discussed in section 3.2.2, *critique* has been taken as an essential part in representing SA as an epistemic practice, recently, some researchers have shifted their attention to assess *critique* explicitly.

Kelly and Takao (2002) proposed a model for university (in California) students' written argument assessment considering epistemic status of the propositions in their arguments in the context of oceanography. In their framework, the epistemic levels are based on discipline specific constructs from lower epistemic levels in which statements are grounded in data to higher epistemic levels where propositions are more general and theoretical. The six epistemic levels are given below (from low to high).

*Table 3.3 Epistemic levels for argumentation analysis (Kelly & Takao, 2002)*

<b>Level 1</b>	Propositions making explicit reference to data charts, representations, locations, and age of islands, or locating the geographical area of study.
<b>Level 2</b>	Propositions identifying and describing topographical features of the geological structure specific to the geographical area of study
<b>Level 3</b>	Propositions describing relative geographical relations amongst geological structures specific to the geographical area of study.
<b>Level 4</b>	Propositions presenting geological theoretical claims or model illustrated with data specific to geographical area of study
<b>Level 5</b>	Propositions in the form of geological theoretical claims or models specific to the area of study
<b>Level 6</b>	General propositions describing geological processes and referencing definitions, subject-matter experts, and textbooks. The knowledge represented may not necessarily refer to data that is specific to the area of study

Their model and analysis method contribute to the methodology of argumentation assessment, but the assessing process is not easy to operate. The model has been found to be limited in assessing whether the propositions are scientifically accurate (Sampson & Clark, 2006). In addition, the model is very discipline specific, which makes it less applicable in other domains. Moreover, the understanding of epistemic level in their study is mainly focused on discipline knowledge without considering the nature of argumentation itself.

In contrast, Sandoval (2003) investigated scientific explanation by not only looking at the disciplinary knowledge involved, but also emphasizing “what a good scientific explanation looks like” (p. 6). He explored high school students' conceptual and epistemological understanding through their explanations of complex events in the context of ‘natural selection’. He referred to the meaning of epistemology as “students' idea of what scientific theories and explanations are, how they are generated and how they are evaluated as knowledge claims” (Sandoval, 2003, p. 8). His study emphasizes two criteria for scientific explanations, one of

which is the *coherent articulation of causal claims (causal coherence)*, which embodies two epistemic goals: (a) they articulate causal mechanisms to explain phenomena and (b) chains of causes and their effects cohere sensibly. The second criterion is *using evidence to support or refute claims (evidentiary support)* which also contains two epistemic goals: (a) cite data explicitly and sense of epistemic criterion of evidentiary support and (b) whether sufficient and appropriate data was cited.

Although Sandoval (2003) is concerned with scientific explanation rather than argumentation, his emphasis on the use of evidence and the connection between evidence and claim implies that explanation is embodied in SA. His study thus provides inspirations for this study in terms of **1) the process of constructing an argument entail explaining how evidence connects to claim, and 2) providing the criteria for evaluating this connection.** However, both Sandoval (2003) and Kelly and Takao (2002) investigated students' epistemic status by looking at the product (explanation and argumentation) students generate, without letting students apply their epistemic understanding directly to evaluate a scientific practice, as this study intended (see Figure 3.1).

Instead of looking at students' epistemic understanding by assessing the arguments they generate, Lee et al. (2014) made things straightforward by asking high school students to identify the uncertainty in an argument and to provide a rationale for that uncertainty. They proposed an assessment framework for SA in their assessment of 'uncertainty-infused SA' in earth sciences, involving claim, justification, uncertainty rating and uncertainty rationale. Uncertainty in science does not only imply that scientific knowledge changes once there emerges new evidence, but that sometimes science conclusions have limitations due to the methods used to investigate them. In other words, there may exist conceptual or empirical errors embedded in investigations. Thus, their study was actually exploring students' epistemic understanding of SA, although they did not say this explicitly.

*Table 3.4 Lee et al.'s (2014) assessment framework for SA*

	<b>Description of the level</b>	<b>Item design in the study</b>
<b>Level 0</b>	Non-scientific	None
<b>Level 1</b>	Scientific claim	Claim
<b>Level 2</b>	Coordination between claim and evidence	Justification
<b>Level 3</b>	Reasoned coordination between claim and evidence	Justification
<b>Level 4</b>	Modified, reasoned coordination between claim and evidence	Uncertainty rating
<b>Level 5</b>	Conditional, modified, reasoned coordination between claim and evidence	Uncertainty rationale

For *Justification* items, they gave scores of 1 to ‘irrelevant’ justifications that are without science-relevant information, scores of 2 to ‘relevant’ justifications with relevant data but without explaining why the data support the claim, scores of 3 to ‘single warrant’ explanations that coordinate between a piece of knowledge and evidence, and scores of 4 to ‘two or more warrants’ explanations that provide more links between evidence and knowledge. *Uncertainty rating* is about how certain students are about their answers, ranging from ‘Not at all certain’ to ‘Very certain’. *Uncertainty rationale* is the reason behind the uncertainty, score 1 represents uncertainty originated from whether and how students’ knowledge, ability, and skill might influence their performance. Score 2 for a rationale that involves the outcome, knowledge, and data related to the scientific investigation featured in the item set. Score 3 represents a rationale that is beyond the investigation featured in the item set.

Lee et al.’s (2014) framework offers insights for 1) evaluating justification and 2) assessing students’ epistemic understanding of SA by asking them to identify (and explain) the uncertainty in SA. However, as they reported, due to their participants’ misunderstanding of *uncertainty*, they didn’t manage to distinguish between the uncertainty caused by students’ uncertainty about their own content knowledge and the uncertainty that exists in the logic of an argument. Their study thus further indicates that **providing enough instructions when assessing epistemic understanding could help ensure what is being assessed is what is aimed to assess**, especially when students know little about the practice.

### 3.3.4 Exploring SA as a learning progression

As mentioned previously, the various analytical frameworks for SA make it challenging for studies to be comparable and to capture students’ development in SA engagement. Researchers realized the need to go beyond nuanced and diverse analysis to make the assessment of SA more comprehensive and available for teaching instructions (Rapanta et al., 2013; Osborne et al., 2016; Henderson et al., 2018). Thus, studies that explore SA as a learning progression have started to emerge in the last ten years. *Learning progression* represents “successively more sophisticated ways of reasoning within a content domain that follow one another as students learn” (Smith et al., 2006, p. 1). It is not a new idea in the field of education but its rise in science education has stemmed from the advocacy of its combination with assessment, aiming to track students’ progress and to align assessment with instructions (Duncan & Hmelo-Silver, 2009).

Berland and McNeill (2010) framed a three-dimensional learning progression that includes instructional context, argument product and argument process, and described how each dimension developed from simple to complex. They justified the framework by analysing the discourse from four classrooms from elementary school to high school in the US and argued that the learning progression is not age dependent but demonstrates the increasing complexity of students' argumentation. Nevertheless, they did not then assess students' argumentation further using their proposed learning progression.

Osborne et al. (2016) constructed and validated a learning progression of SA by considering argumentation as “a process of construction and critique” (p. 825). Their study was conducted with middle school students in the US and the instrument is contextualized in the physical behaviour of matter. Their learning progression consists of three main levels, and several sublevels that are described within each main level (see Table 3.5). The three broad levels of the learning progression were well validated. They also reported that the students benefited from scaffoldings in items (in the form of sentence starters), which helped them engage in critique and complete the argumentation task. The assessment items they designed include only open-ended questions due to the poor performance of the multiple-choice questions in their initial item pool. At the end of their study, they mentioned that the sublevels of the learning progression need further validation as they were not revealed in the students' performance data; it was also unclear whether the learning progression would change in other cultures such as in China and this needs to be explored as well.

*Table 3.5 Osborne et al. 's (2016) learning progression*

<b>Level</b>	<b>Constructing</b>	<b>Critiquing</b>
<b>0a</b>	Constructing a claim	
<b>0b</b>		Identifying a claim
<b>0c</b>	Providing evidence	
<b>0d</b>		Identifying evidence
<b>1a</b>	Constructing a warrant	
<b>1b</b>		Identifying a warrant
<b>1c</b>	Constructing a complete argument	
<b>1d</b>	Providing an alternative counter argument	
<b>2a</b>	Providing a counter-critique	
<b>2b</b>	Constructing a one-sided comparative argument	
<b>2c</b>	Constructing a two-sided comparative argument	
<b>2d</b>	Constructing a counter claim with justification	

Osborne et al. (2016) is a significant study because it provided empirical evidence obtained from a large-scale assessment for the possibility of exploring SA as a learning progression. The

SA learning progression informs the understanding of the nature of SA in terms of how it develops and becomes more sophisticated (Osborne et al., 2016), which further enlightens how teaching/instruction can be operationalized. Thus, their study sheds light on **the possibility of exploring how the competences needed for SA engagement may develop**. The assessment instrument developed also provides insights for developing pencil and paper SA assessments that can be used for large-scale administration.

### **3.3.5 Guidance for developing SA assessments**

As revealed from the discussion in the previous sections, analytical frameworks for SA are many but assessments that can be used for large scale administration remain scarce. Thus, it becomes a problem of how SA assessments can be designed. This section will review literature, that may not talk about SA assessment specifically, but that can inform SA assessment design.

Berland and McNeill's (2010) SA learning progression considers the instructional context in which SA happens, including whether the question is closely defined or has multiple potential answers, the size of data set required, the appropriateness of data in the data set, and the extent of scaffold provided. Their study provides insights for designing SA assessments since instructional context is part of a progression and part of assessment design (Duncan & Hmelo-Silver, 2009; Shute et al., 2016). While Berland and McNeill (2010) asserts that items including both appropriate and inappropriate data are more complex than those including only appropriate data, Ahmed and Pollitt (2001) found in a science test that irrelevant information distracts students' attention especially during exams when they are under stress. Likewise, Crisp et al. (2008) examined students' expectations of a science test and found that the students expect every piece of information in a test item to be useful, even when it isn't.

Crisp et al.'s (2008) study explored the influence of students' expectations for a science test in terms of validity, and categorized the mismatch in expectations into three levels, namely, 'subject level' (the assessed construct and how it is assessed), 'question level' (irrelevant information and item difficulty), and 'sentence level' (language). They reported how changing the wording of items reduced the threat to test validity. Similarly, Ahmed and Pollitt (2001) argued that language in a real-world context is usually more complicated than that in context-free scientific scenarios thus reading ability is often needed for understanding the assessment questions.

Considerations of scaffolding as described in Berland and McNeill (2010) is discussed in other

studies as test-takers' familiarity with what is assessed. What test-takers think they are being assessed on can easily mismatch with what the test designer intends to assess especially when test-takers are unfamiliar with the assessment type/content (Cheng & DeLuca, 2011; Crisp et al., 2008). Ahmed and Pollitt (2000) found that students tend to think that a test is assessing new scientific knowledge when in fact it is often assessing familiar knowledge in a new context. Therefore, some amount of instruction for the test-takers that helps them familiarize themselves with the assessment can improve the validity of test scores (Koretz et al., 2001). Deane et al. (2019) talks about the advantages of using a scenario-based assessment to assess written argument in reading and writing. They found that the students who first finished lead-in tasks expressed their ideas more efficiently in the final essay compared with those who first finished the essay, and the former task sequence reduced their cognitive load thus focusing their attention on the process of writing. They further asserted that lead-in tasks and the task sequence supported the students' writing and offered instructional information to them. However, a potential problem often present in scenario-based assessments is item dependence, which occurs when different items use a common scenario (Wang et al., 2005). Despite the importance of familiarity asserted by these studies, Haladyna and Rodriguez (2013) argued that new materials should be used to elicit higher-level thinking and avoid assessing recall/recognition.

Ng Yee Ping (2019) explicitly talked about designing SA assessment items and proposed a 'Three-cornerstones' model for designing SA items, that is, 'argumentation', 'item anatomy', and 'learning objectives'. The author argues that designing an SA assessment should consider the argumentation skill, item features, and scientific knowledge embodied within it. Ng Yee Ping (2019) provided several points for each cornerstone to consider by summarizing previous research around topics such as the item format, test length, and scenario selection in 'item anatomy'. Haladyna and Rodriguez (2013) asserted that item formats that allow for extended writing or articulation of reasoning are good at assessing higher order thinking skills. Other authors however have criticized constructed-response assessments for not detecting the process and components that contribute to the response and for the fact that a single score provides limited information (Deane et al., 2019).

Overall, there are few studies discussing how to design assessments for science learning, and even fewer that explore how to design assessments for higher-order thinking skills such as SA (Shute et al., 2016). Research that provides empirical evidence for SA assessment design is

therefore needed to advance the research and practice of SA assessment.

### 3.3.6 Assessment format for SA

Previous studies exploring the assessment of higher-order thinking skills and scientific practices have tended to adopt observation or open-ended questions to collect data. Examples of studies drawing on observations include Chen and Terada (2021) and Shi et al. (2021), which used video-recorded lessons to analyse students' scientific practice and high school teacher's teaching performance respectively. Similarly, Erduran et al. (2004) approached the analysis of SA by analysing audiotaped lessons. As for written forms of assessment, most studies especially large-scale studies used tests comprising of open-ended items to assess SA or other competences (Osborne et al., 2016; Dawson & Carson, 2017; Wang & Song, 2021). There were few studies that using tests that include close-ended items when assessing competences. For instance, Romine et al. (2017) designed the QuASSR that contains close-ended items to assess student's social scientific reasoning ability, although they assigned a score of 0-2 for choosing each of the options in an item.

This study aimed to construct a broad picture of Chinese high school students' SAC and thus large-scale data were required. Therefore, a test was appropriate for this study although observations and essays, which are suitable for small scale study, have been used to analyse SA. In the case of large-scale SA assessment, Lee et al. (2014) used online test, which has been taken as a low-cost and time saving way to assess SA (Nardi, 2014). But pencil and paper tests were considered as appropriate for this study as taking online tests was not a common practice for Chinese high school students since they usually take pencil and paper tests, the accessibility of computers for the students might be a challenge, and the impact of computer-based tests may have on their test experience could be an issue (Cheng & DeLuca, 2011). Additionally, given there was no existing tests that can reflect the construction of SAC of this study and can be used in the Chinese context, this study intended to design a test. Despite the wide use of open-ended items to assess SA, close-ended items were also designed in this research for its practicality in large-scale tests, e.g., scored reliably and cheaply (Osborne et al., 2016).

Several enlightenments for carrying out this research can be obtained by the discussion in the previous sections. Firstly, existing studies only investigated SA assessment/analysis from one or two aspects of the *structure*, *content*, and *epistemic understanding* of SA. Exploring a comprehensive way to assess SA is thus worth considered. Secondly, content knowledge



should be considered in the scientific context for assessing argumentation. Thirdly, students' epistemic understanding of SA can be assessed but needs careful consideration of students' understanding of assessment items to obtain valid responses from them. Fourthly, it is plausible to consider exploring SAC as a learning progression and worthwhile to see how SA learning progressions may be different when assessing students in different cultures/countries. Lastly, not only the SA skill entailed in an item, but item design also influences the complexity of an item, and empirical evidence in terms of how SA assessment items can be designed is needed.

### **3.4 Students' SA engagement**

This chapter has talked a lot about SA from a theoretical perspective so far, but how does SA typically manifest in students' engagement? The ability of engaging in argumentation can emerge from an early age and develop with age naturally, but deliberate instructions are needed for some aspects of it and for it to be skilful (Kuhn et al., 2010; Rapanta et al., 2013). This section will discuss the aspects of SA that are more difficult/easier to engage in and what has been found to influence students' SA engagement. The discussion focuses on argumentation as a result of instruction rather than age, therefore providing information for understanding students' development of SAC with a focus on the aspects that need more instructions and for designing appropriate assessment items to elicit SA.

Students often have existing views about scientific practice like SA, and it is of value to explore these views since they influence their engagement in the practice (Hudicourt-Barnes, 2003; Bricker & Bell, 2007; Bricker & Bell, 2008). McNeil (2011) conducted a study exploring how 5<sup>th</sup> grade students in the US view argumentation before and after a school year's course that provides instructions for argumentation. They found that students focused more on 'disagreement' when talking about the argumentation among scientists, but they tended to have no idea of its meaning in science classrooms, and it is in the everyday life setting that they mentioned more about 'emotional or angry fighting'. But at the end of the school year, the students' views did not show much difference across the three contexts (i.e., among scientists, in classrooms, and in everyday life), with focus shifted to 'exchange between people'. Similar findings were reported in Bricker and Bell (2012) that middle school students tended to associate 'argument' with yelling and fighting when they had not provided them with a specific context to understand the term. But there is a scarcity of evidence indicating how students in China view SA and how older students perceive 'argument' differently than primary and middle school students.

Studies conducted both in the US and in China have reported that *claim* is the element students are most likely to identify in others' argument and the easiest to construct in their own argument, and it is easy for students to use *evidence* in their writing (McNeill et al., 2006; Deng & Wang, 2017; Berland & Reiser, 2009). However, connecting between claim and evidence and articulating why a piece of evidence supports the claim, namely, generating *reason* is a more difficult aspect of SA (Deng & Wang, 2017; McNeill & Krajcik, 2007). Moreover, Berland and Reiser (2009) examined middle school students' weaknesses and strengths of engaging in SA in the US and found that students had difficulties in differentiating between claim, evidence, and reason, and failed to articulate them in an argument rather by weaving these elements together. Similar findings were found in Sadler's (2006) study in which preservice teachers in a university in the US struggled with identifying warrants from evidence. The above concerns the justification part of SA, as for the critique aspect, studies revealed that students have difficulties and are unaware of attending to other's arguments and tend to focus on supporting their own claim when there are multiple claims in a conversation (Chen et al., 2019; Kuhn et al., 2010; Kuhn & Udell, 2007).

To reach a more comprehensive understanding about SA and therefore its assessment, the following subsections will review research evidence on factors associated with SA performance. The discussion will focus on how SA relates to content knowledge, instruction, and the cultural context, because these factors are related to the present study.

### **3.4.1 SA and content knowledge**

The relationship between content knowledge and scientific argumentation has always been a focus in argumentation research. Given the aim of this study is about SA assessment and exploring students' SAC, this section will mainly talk about how students' content knowledge could affect their argumentation rather than how engaging in SA could influence content knowledge proficiency. There are several questions here, for example, could the ability in terms of argumentation be separated from the level of content knowledge? Would students with higher levels of content knowledge perform better in argumentation? Would it be more difficult for students to engage in SA when more content knowledge is needed in the specific question? Figuring out the relationship will contribute to a deeper understanding of SA and therefore its assessment.

Based on the nature of SA as discussed previously, *firstly*, when talking about argumentation

per se, whether taking argumentation as a process or a product, people must argue about something, people cannot produce an argument or conduct an argumentation without arguing about something. Obviously, people cannot engage in argumentation without knowledge, no matter what kind of knowledge they will use. So, argumentation must be conducted under the context of one or several domains, as many researchers have argued (Toulmin, 1958; Mcpeck 1981; Govier, 2018). *Secondly*, competence, also known as skill or ability, is a dispositional concept (Fischer et al., 2014). According to Heider (1958), if someone has a certain skill, he or she will manage to do the thing appropriately and successfully when given the motivation and opportunity to do so. From this angle, people care more about if a person argues well in one domain, can he or she also perform well in all other domains, and if it is possible to teach students argumentation skills that can be used across domains? Thus, some domain-general characteristics of certain competence are desired. Theoretically, it seems that content knowledge is a necessary but insufficient condition for engaging in SA.

Empirical studies have shown that students tend to engage in argumentation when they are familiar with the topic, and students with high prior knowledge outperformed low prior knowledge students (Von Aufschnaiter et al, 2008). Yang et al.'s (2015) study in four 8<sup>th</sup> grade science classes showed that high prior-knowledge students significantly outperformed low prior-knowledge students. However, Sadler and Donnelly (2006) conducted a study in an urban high school in the US that suggested that content knowledge alone does not necessarily result in improved argumentation. They conducted both qualitative and quantitative analysis that indicated that content knowledge was not a significant factor of argumentation quality. Since their study was conducted under the context of socio-scientific topics, they made several explanations that may account for the results. One of the interpretations is a threshold model, which refers to the non-linear relationship between knowledge and argumentation that students need to have a certain degree of basic knowledge before the level of knowledge can affect their argumentation performance (Sadler & Donnelly, 2006). This assertion is commensurate with the notion that some general skills require a certain minimal amount of content knowledge which they can operate on (Alexander & Judy, 1988; Fischer et al., 2014). Wang and Buck (2015) modified Sadler and Donnelly's (2006) model by examining the relationship between Chinese middle school students' SMK (Subject-matter knowledge) achievement and argumentation engagement. Interestingly, they found that medium-SMK students showed better understanding of taking SA as a knowledge construction process and had greater potential in argumentation than low-SMK and high-SMK students, although sometimes they

tended to cite more inaccurate knowledge in their arguments. Osborne et al. (2016) investigated the assessment of SA and found that items under scientific context have higher difficulty estimates than items under social scientific context. However, there is little empirical evidence to show how exactly and to what extent prior knowledge influences students' performance on argumentation, but the core idea is that students cannot argue without a certain level of scientific knowledge.

There are also empirical studies providing evidence to support the existence of domain-general factors in SA. For example, Zohar and Nemet's (2002) study found that 9<sup>th</sup> grade students can transfer their argumentation ability in the context of genetics to other daily life context. However, Siler and Klahr (2016) conducted a study that showed that 6<sup>th</sup> -7<sup>th</sup> grade students engaging in argumentation under domain-specific context and using concrete information could better transfer the skill to another domain than those who engaging in abstract context.

Taken together, it is impossible to argue without any knowledge, but a higher level of knowledge does not necessarily indicate better performance on argumentation. Put simply, there are some aspects of argumentation which could transfer across domains while others are specific within a domain and even to a topic. This gives this study two implications. Firstly, it is inevitable that we need to consider students' accuracy of knowledge when analysing their argumentation. In addition, it is important to make the assessment appropriate to students' understanding of the related knowledge, especially that the content knowledge should not be too difficult or unfamiliar for them. Secondly, it is valuable to pay attention to domain-general factors when understanding and assessing SA. The latter point will be discussed in the next section that talks about how general instruction on argumentation influences SA engagement.

### **3.4.2 SA and instruction**

The 'instruction' here aims to deal with the question of 'how much students know about SA and how would that influence their engagement in SA?'. It has been reported widely that SA can be advanced by explicit instruction. The instruction discussed here is not about content knowledge, thus studies reviewed in this section can also shed light on how domain-general knowledge influences SA performance as mentioned in the previous section. In addition, it informs the understanding of how epistemic understanding of SA influences SA engagement, as mentioned in section 3.2.2.

Zohar and Nemet (2002) examined how a 12-hr unit that teaches SA explicitly in the context

of dilemmas in human genetics affected ninth-grade students' performance on biology knowledge and SA in Israel. In the unit, they explained the definition and structure of SA and the criteria of distinguishing between good and bad SA. In addition, the students were given opportunities to practice engaging in SA in certain content contexts. Their study conveyed quite positive messages about the usefulness of instructions on SA performance. They found that the students became more aware of the importance of specific content knowledge for constructing argumentation and tended to construct more complex written arguments with more justification after the unit. Moreover, the students also demonstrated higher quality of SA in the discussion, with more explicitly expressed conclusions and more justifications.

However, it is too early to say that instructions can always bring positive results on students' SA engagement. Sadler (2006) investigated preservice science teachers' perceptions about SA and their ability of forming and evaluating arguments during a six weeks' intervention course in a U.S. university. During the intervention, he spent two lessons providing some explicit instructions about SA, including identifying argument structure based on simplified TAP and how to improve argument effectiveness by considering counter-positions and rebuttals. The mixed finding of his study shows that some participants formulated more complex arguments after the explicit instructions. However, several participants showed no improvement, of which most demonstrated very low-level performance before the instruction. Interestingly, some participants showed improvement right after the instruction but reduced to the level before instruction at the end of the course. His study demonstrates that to make instruction effective, both the **time** and the **type** of instruction should be considered. Another point to note is that both Zohar and Nemet (2002) and Sadler's (2006) study considered only the structure aspect of SA when assessing it, namely, assessing SA based on the number of functional statements contained within it.

As previously mentioned, the focus of research in SA has been shifting from the structure of SA to take it as an epistemic practice in which evaluation plays a significant role. Lombardi et al. (2018) adapted a scaffold, namely Model-Evidence Link (MEL) diagrams and Model-Evidence Link Tables (MET), that helps with students' evaluations about the connections between multiple lines of evidence and alternative explanations about a phenomenon. They investigated the difference of the influence of these two scaffolds on high school (in the U.S.) students' evaluations and knowledge construction with another scaffold (Mono-MEL) that does not contain alternative explanations about a phenomenon. Specifically, the students

participated in the scaffold-based instructional activities including four topics in earth science over the course of a single school year. They found that the scaffold that helped students evaluate alternative explanations were more effective in facilitating the students' evaluation and knowledge construction, and MEL performed even better than MET.

The above intervention studies provide information in terms of whether and to what extent instruction may improve students' performance on SA. Another point related to instruction is how ignoring it may prevent researchers from getting authentic information on the students' SA performance. This is usually mentioned when studies aim to assess students' SA performance in cases that the students have never encountered the term/activity explicitly, but they may have the ability implicitly. For instance, Osborne et al. (2016) mentioned how providing sentence starters in their assessment tasks helped the students be aware of the requirement of tasks.

To sum up, knowing/not knowing the definition and structure of SA can influence students' SA performance. The mixed and sometimes-contradictory findings could suggest that the time length of instruction, types of scaffolds, and the context in which it is applied can have different effects on participants' SA engagement. So, taking the scaffold into consideration when conducting the assessment can help reveal whether and to what extent the students' poor SA performance is due to lack of awareness of the activity or lack of competence. It can also inform the ways in which to equip students with SAC.

### **3.4.3 SA and cultural context**

It has been widely recognized that people with different social and cultural background tend to have different values on discourses where conflict is potentially involved, and that will influence their performance in argumentation (Kuhn et al., 2010; Schwarz & Baker, 2016). Researchers have argued that argumentation is not a normal feature of East Asian discourse in which insisting on different opinions are often regarded as being in a position of personal rivalry and compromise solutions are appreciated (Osborne et al., 2016; Becker, 1986).

However, studies that investigate how argumentation connects to cultural background have mixed findings. Kuhn et al. (2010) compared how students (aged around 12 years old) from China and the US perceived the value and goal of argumentative discourse using two scenarios. One scenario provides opportunity for them to optionally engage in social interaction. Whereas it is a necessity to engage in interaction for the other scenario. They found that the students

from the US were more willing to interact with their peers than the Chinese students in the first scenario. But in the second scenario, Chinese students were more likely to choose to engage in interaction. However, the scenarios in their study are not in scientific contexts. So, their study revealed a more general picture of how cultural context might influence students' willingness of engaging in discussions that contain potential conflict. Similar findings were reported by Dong et al. (2008) that investigated the collaborative reasoning of 4<sup>th</sup> grade students in China, Korea, and the US. They found that Chinese students and Korean students can easily adapt to collaborative reasoning and showed high engagement, and the pattern of their argument stratagems were similar with those of American students. However, Jin et al. (2016) found that Chinese students showed lower levels of argumentation than explanation and U.S. students were the opposite when dealing with a social scientific issue.

Studies have provided information that is contrary to those who claim that Asian culture hinders students' performance in social discourse. Zeidler et al. (2013) examined the difference of socio-scientific issue (SSI) reasoning and epistemological beliefs between high school students in five regions (i.e., Jamaica, South Africa, Sweden, Taiwan, and USA). Their finding did not reveal much difference between these five areas, except for the Taiwanese students. The Taiwanese students in general demonstrated higher degrees of epistemological sophistication in their justifications and questions about the SSI issue; their views about the nature of knowing and the structure of knowledge in science indicate that they view science as "an interrelated network of highly integrated concepts" rather than "an accumulation of discrete facts" in isolation (p. 274). But the contradictory finding of this study could be due to its focus on justification, where students were not exposed to competing claims. Framed arguments in a more cooperative orientation rather than display identity or assert dominance, Chinese students were found less pessimistic about the consequences of conflicts and didn't show more avoidant of confrontation and interpersonal argumentation (Xie et al., 2015). Other studies revealed that engaging in SA facilitated Chinese students' engagement in Physics class (Wang & Bunk, 2015) and their SA performance (Shi, 2019, 2020; Luo et al., 2020).

Combining the above studies, cultural context indeed seems to demonstrate effects on students' performance of engaging in social discourse especially argumentative discourse. In their investigation of a learning progression of SA, Osborne et al. (2016) also mentioned that their progression is based on data of students with a Western cultural background and the performance of students in other cultures might be different. It is therefore important to conduct

SA assessment research in China to inform SA teaching, learning and assessment in China.

However, these studies do not explore how Chinese students perceive argumentation and whether they are willing to engage in argumentation despite their capability of doing it. Students' understanding about the nature, purpose and value of activities in which individuals hold different opinions is likely to influence their engagement and its productivity (Kuhn et al., 2010; Schwarz & Baker, 2016). Among the few studies that exploring how students perceive SA, Heitmann et al. (2017) found that German students in secondary school tended to perceive the role of facts rather than discursive characteristics as highly relevant for science lessons; Kaya et al. (2010, 2012) reported that high school students in Turkey felt enthusiastic about engaging in SA (similar findings were reported by Cetin (2014)) and positioned SA across various aspects of science learning such as 'understanding rather than memorizing', 'getting new ideas', 'understanding the nature of science' etc. Recently, Ke et al. (2020) explored U.S high school students' perception of SSI-based learning and found that students in general appreciated the learning experience and found SSI engagement relevant, interesting, promoting agency, and beneficial for their science learning. Bathgate et al. (2015) found that there was a significant positive relationship between students' being willing and being able to participate in SA, but students who held a negative value towards SA gained significantly fewer benefits from engaging in SA.

Overall, it remains a question whether Chinese students' capability of engaging in argumentation is a 'compromise' to 'being required', given that East Asian students were found more likely to show inconsistency in attitudes and behaviour (Spencer-Rodgers et al., 2010). Additionally, with the deepening of globalization over the past decades, it needs to be reconsidered that whether or to what extent students who may still be influenced by the naïve dialecticism in Eastern culture appreciate, capable of, and benefit from argumentation.

### **3.5 Assessment validation**

The previous review lends insight into understanding SA generally and as encountered by untrained students thereby informing the framing of SAC and the design of a SAC assessment. However, assessment score interpretation and use need to be justified by empirical evidence derived from scientific inquiry together with rational argument (Messick, 1995). Thus, validating an assessment is an essential part of assessment development and transparent documentation of validation allows an assessment to be critiqued, adapted, and thus improved.



This section will discuss studies that contribute to assessment validation to inform validating the SAC assessment in this study. Given the debates in the field of validation on the meaning of ‘validity’ and ‘validation’, this section begins by explicating the understanding about the terms in this study. Firstly, we are talking about validity based on the modern validity theory in which validity has one unifying conception, namely construct validity (Lane et al., 2015). Secondly, ‘validity’ is not a property of a test itself but a property of the interpretation of test results, and it is the interpretations or uses of test results that need to be validated (Kane, 2016). Thirdly, ‘validation’ is an ongoing process to evaluate a proposed interpretation of the test results, relying on multiple evidence sources which might be different based on different interpretations (Shaw & Crisp, 2012; Kane, 2016). Based on these basic understanding, this section introduces some perspectives and practices about how to conduct validation in assessment.

### 3.5.1 Argument-based approach for validation

As mentioned previously, there has been a consensus that validation evidence should be from multiple sources and be connected to support the proposed interpretation of the test results. Therefore, the need to organize evidence into a persuasive argument rather than only listing supportive evidence has been recognized in the field of validation. Interestingly, most studies that propose approaches for conducting validation research are based on Toulmin’s argument pattern (TAP) (Newton, 2017).

One of the most influential approaches to validation, the argument-based approach, was proposed by a series of Kane’s (2009, 2012) studies. His approach is concerned with two main points for validation: what is being claimed and how to support the claim. Thus, his approach includes two steps. **One** is to specify the interpretation and use of an assessment by articulating the “network of inferences and assumptions leading from test performances to conclusions and decisions based on the test scores” (Kane, 2012, p. 8). By doing this, an **interpretation/use argument (IUA)** is constructed, which is an argument about the supposed interpretation/use of an assessment. The IUA as a whole and the inferences and assumptions in it can be taken as a network of claims that need to be supported by empirical evidence or logical analysis. **The other** step is to evaluate (support or falsify) the IUA critically using evidence from multiple sources and various analysis. By doing this, a **validity argument** is constructed. Kane (2006, 2009) proposed three main inferences, and Table 3.6 below shows what an IUA looks like. Warrant here supports the inference, and the assumptions need to be supported to justify the

warrant. An IUA serves as a framework to guide a validation, in other words, the empirical evidence and logical analysis are organized based on the IUA to form a validity argument. Kane (2016) pointed out that there is no single pattern of an IUA, and the inferences and assumptions are not a checklist rather they are examples. He also claimed that the construction of the IUA depends on and should be aligned with the interpretation/use of test results.

*Table 3.6 IUA proposed by Kane (2006, 2009)*

<b>Inference</b>	<b>Warrant</b>	<b>Assumption example</b>
<b>Scoring</b>	Test score is a faithful representation of performance on a particular test.	Scoring rules and procedure is appropriate. Scoring rules are consistently and accurately applied.
<b>Generalization</b>	Test score can be taken as universe score that represents universe performance.	Test tasks/items are representative for the targeted performance.
<b>Extrapolation</b>	Test score can reflect a wider performance in the domain.	Test tasks/items are representative for the targeted domain.

As an extension of Kane’s (2006, 2009) framework, Shaw and Crisp (2012) added two other inferences to their interpretation argument. Table 3.7 illustrates their framework for the interpretation argument.

*Table 3.7 Interpretation argument proposed by Shaw and Crisp (2012)*

<b>Inference</b>	<b>Warrant</b>	<b>Assumption example</b>
<b>Construct representation</b> (Task-test performance)	Tasks elicit performances that represent the intended constructs	Constructs can be identified. It is possible to design tasks that require these constructs.
<b>Scoring</b> (Test performance-test score)	Scores reflect the quality of performances on the assessment tasks	Rules, guidance, and procedures for scoring responses are appropriate for providing evidence of intended constructs. Rules for scoring responses are consistently and accurately applied.
<b>Generalization</b> (Test score-test competence)	Scores reflect likely performance on all possible relevant tasks	Enough tasks are included in the test to provide stable estimates of test performances.
<b>Extrapolation</b> (Test competence-domain competence)	Scores reflect likely wider performance in the domain	Constructs assessed are relevant to the wider subject domain beyond the qualification syllabus.
<b>Decision-making</b> (Domain competence-trait competence)	Appropriate uses of scores are clear	The meaning of test scores is clearly interpretable by stakeholders who have a legitimate interest in the use of those scores i.e., admissions officers, test takers, teachers, employers

As shown in Table 3.6 and Table 3.7, both frameworks emphasize more on the product of assessment and the psychometric quality of it (i.e., mainly psychometric evidence is required),

although Shaw and Crisp (2012) added a ‘construct representation’ inference. These two approaches care more about what happens after the test has been designed and delivered, thereby the IUA they constructed do not elaborate the process of developing an assessment.

On the contrary, some researchers that embrace constructing validation using an argument-based approach recognize the importance of including claims about the whole assessment process into the IUA of its validation. Examples include the step-based lifecycle model mentioned by Newton (2017), and the validity-by-design perspective of Mislevy (2007). Among these researchers, Ferrara and Lai (2015) provided a detailed framework for constructing an IUA in which claims are made and evidence should be collected throughout the assessment process. They specified seven steps in their framework to help construct a validation argument:

- 1) determination of testing program policies and articulation of intended interpretations and uses of test scores,
- 2) test design and development,
- 3) test implementation,
- 4) response scoring,
- 5) technical analysis,
- 6) delivery of scores and other feedback to test users, and
- 7) interpretation of score reports to guide decisions and take other actions.

Under each of these steps, they provide the claim that needs to be justified and the evidence that can be used. Detailed information about their framework can be found in Appendix 2.

### **3.5.2 The macro- and micro-validation perspectives**

Recognizing the distinctions of the different approaches of conducting validation research, Newton (2017) proposed a macro-micro validation continuum in which validation can be approached from two perspectives. ‘Macro-validation’ foregrounds the product-related questions of assessment, such as the testing results, uses and consequences of an assessment. It tends to deal with the overarching claim (such as ‘it is possible to measure the target proficiency accurately using assessment results’) directly. Evidence in a macro-validation investigation is usually to justify the outcomes of an assessment procedure as a whole, such as the relation between the assessment results and other variables. Thus, macro-validation usually begins after the assessment results have been delivered.

In contrast ‘micro-validation’ underlines the procedure-related questions of the assessment and aims to validate the effectiveness of the features and processes that constitute the assessment procedure. Once an assessment design starts, the micro-validation begins. Evidence such as the analysis of test-taker’s response process is close to the micro end of validation. However, it is the use to which the evidence is put that determines the perspective it should be pertained to, not the type of evidence (Newton, 2017). Newton (2016) claims that if the process of producing a product is well controlled, the product is very likely to be of high quality. Thus, micro-validation and macro-validation complement each other in reaching a coherent validation argument.

Newton (2017) did not provide any specific framework for he claimed that:

“...any source of evidence or analysis that helps to establish a case for or against the overarching measurement claim (that it is possible to measure the target proficiency accurately using assessment results) should be considered a legitimate source; whether or not it seems to fit neatly within any of the established frameworks.” (p. 42)

Taking the above approaches and perspectives on conducting validation study together, it can be concluded that:

- 1) it is necessary to consider validation as the construction of an argument in which various evidence are used to support a network of claims formed by a variety of inferences and assumptions.
- 2) validating the effective design of an assessment procedure is an important part of validating the interpretation/use of assessment results.

### **3.5.3 Including test-takers’ voices into validation**

Section 3.1 has talked about the importance of listening to students due to the impact assessments always have. The importance of including test takers’ voices as one of the sources of validation evidence has also been recognized by researchers in the field of educational assessment. But the fact is that students’ voices are more often ignored in assessment studies (Cheng & Deluca, 2011). This section will discuss some studies/perspectives on considering students’ voices when developing and validating an assessment to justify the plausibility of exploring students’ test taking experience in this study.

Investigating students’ response processes on assessment items contributes to a deeper understanding of whether the items elicit the skills that are targeted by the assessment. This is

not a new application of test-taker's voices, as it has been mentioned in many frameworks in terms of assessment construction and validation (Kane, 2013; Wilson, 2004; Newton, 2016). But it is still worth illustrating it here for few studies report how students' voices contribute to their work in an explicit and transparent way. Cohen and Upton (2006) conducted a study to investigate the performance of the reading tasks in the new TOEFL test by exploring the test-taking strategies test-takers use. They found that students did not apply the supposed skill (i.e., academic reading skill) when dealing with the new task rather taking it as another test-taking task. Other studies also uncovered the relationship between psychological factors and the test results. These factors included anxiety (Eklöf and Nyroos 2013), the unfamiliarity with the way of testing (Cohen & Upton, 2006), and dislike with the way of testing (Cohen & Upton, 2006).

Cheng and DeLuca (2011) asked 59 students in a university in Asia to write a short report about their experience of taking English language assessments to explore how test takers' voices can contribute to assessment validation. Students in their study showed that they have their ideas on the test structure and content in terms of whether the test can elicit authentic and reliable performances. Their finding revealed that insufficient time limited students' ability to show their real ability, while a very long test without sufficient breaks resulted in fatigue and low performance on the final items.

To sum up, students' voices are more and more recognized as an important source of evidence for assessment validation, although it is still rare for assessment practice to appeal to students' perspectives for assessment structures, format, and design (Elwood et al., 2017). Thus, it is of value to explore students' experiences and perspectives on an assessment to unearth possible construct-irrelevant variability that might influence the validity of the assessment.

This section talked about how an argument-based approach for validation is understood and operated; how a macro-micro perspective of validation well complemented the discrepancy of various approaches for validation; and how a commonly overlooked source of evidence- 'students' voices'-can contribute to the validation of an assessment. In this study, all the above three approaches will be employed when validating the SAC assessment, which will be illustrated in Chapter 6.

### **Chapter summary**

The review of literature in this chapter highlights the lack and complexity of empirical research exploring large-scale and comprehensive SA assessment to understand SAC, as well as the

scarcity of empirical research exploring Chinese students' SA engagement. This chapter first emphasized the importance of conducting assessment research to improve teaching and learning. By reviewing how SA has been understood as *product, process, individual activity, social activity, as persuasion, as collaboration, and as epistemic practice*, this chapter argued that understanding SA from a competence perspective can help integrate the various perspectives of SA. A three-component framework for SAC including *identifying a scientific argument (I-SA), evaluating a scientific argument (E-SA), producing a scientific argument (P-SA)* was considered possible to frame SAC.

The review of literature on SA assessment/analysis revealed the trend and need to explore SA assessments that are comprehensive and can uncover how the ability develops from less to more skilled. The *structure, content, and epistemic* aspects of SA need to be considered to design a comprehensive assessment, exploring SA as a learning progression is possible and useful for advancing SA understanding and teaching. Toulmin's argument pattern was presented to be a useful framework to understand the structure of SA, and Osborne et al.'s (2016) study was found to be valuable for comparison given it is nearly the only large-scale empirical SA assessment research exploring SA as a learning progression. Some aspects of SA were often found to be more difficult for untrained students to engage, which can be used to inform a possible learning progression in this study. All these considerations need to be evaluated by empirical evidence obtained in praxis. The mixed findings in terms of the relationship between SA and *content knowledge, instructions, and cultural context* revealed the need to understand SA by praxis and under specific context by considering these factors in designing a SAC assessment and understanding Chinese high school students' SAC.

Given the complexity of SA, the lack of research on SA assessment, and the lack of research illustrating how to design SA assessment, it is particularly important to validate the assessment and to make the process of developing a SA assessment transparent and justified. Kane's (2009) argument-based approach for validating assessments was found to be useful because it allows for the transparent documentation of not only evidence but also the logic of justifying the interpretation of the assessment results. Newton's (2017) Macro-micro validation approach and Ferrara and Lai's (2015) validation framework was presented as appropriate to guide the assessment validation in this study because they are concerned with not only the product but also the process of developing an assessment. Lastly, test-takers' voices are worth considered to examine the impact brought by an assessment and obtain validation evidence. Overall, this

chapter has laid the foundations for the value, aim, and design of this study. The next chapter will introduce the philosophical grounds, the research design, and ways of collecting and analysing data in this study.

## **Chapter 4. Methodology**

### **Introduction**

Chapter 1 has provided the overall research aim and research questions, Chapter 2 and 3 has identified the necessity of conducting the research in China and the possible ways of conceptualizing SA and framing the research design of this study. This chapter will provide a detailed explanation of how this study was conducted to address the research aim and answer the research questions.

Specifically, this chapter starts by discussing how a Pragmatism philosophical approach has supported the study. Then the mixed methods design of the research and how it supports addressing the research aim and research questions will be justified. The approach and procedure of sampling, data collection, and data analysis will be elaborated and justified while considering limitations. Finally, the ethical issues encountered during the research will be discussed.

### **4.1 Philosophical underpinning**

Social research is categorized by paradigms that are philosophical assumptions in terms of the nature of reality and knowledge (Lincoln et al., 2011; Kaushik & Walsh, 2019). The term ‘paradigm’ was first introduced by Kuhn (1970), referring to shared generalizations, beliefs, and values of a community of researchers about “which questions are most meaningful and which methods are most appropriate for answering those questions” (Morgan, 2007, p. 53; Kaushik & Walsh, 2019). Each paradigm shares different views in terms of axiology, ontology, epistemology, methodology, and rhetoric of research (Kaushik & Walsh, 2019).

Two extremes along the paradigm continuum are positivism and interpretivism, this is because positivists view our knowledge of the world as objective and researchers should be objective analyst that dissociates from personal values and beliefs, while interpretivists view human’s knowledge of the world as subjective and created by human conceptions (Morgan, 2014a; Žukauskas et al., 2018). Therefore, positivists seek to use quantitative research methods that are not affected by the researchers’ prejudices, while interpretivism is often associated with qualitative research methods to interpret the world. In this metaphysical discussion of the philosophy of knowledge, different assumptions of ontology and epistemology lead to different knowledge that is possible to be generated and ways to obtain it. Thus, the world and ways to



know it are taken as either objective or subjective from the view of traditional philosophy of knowledge.

Instead of holding a dualistic view of the world and involving in the debate of traditional metaphysical philosophy, pragmatists “get over” rather than “solve” traditional philosophical problems by focusing on human *experience* in the world (Morgan, 2014a, p. 1049). Pragmatism has been criticized as “telling us nothing about their ontology or epistemology” (Lincoln, 2010, p. 7) and been reduced to practicality associated with Mixed methods (Morgan, 2014a). However, pragmatism just understand the world by going beyond talking about ontology or epistemology (or discuss philosophy of knowledge from a new perspective), and it is inherently a philosophy despite its practicality on research design (Morgan, 2014a; Kaushik & Walsh, 2019). Figure 4.1 summarizes how pragmatism as a philosophy understands the world.

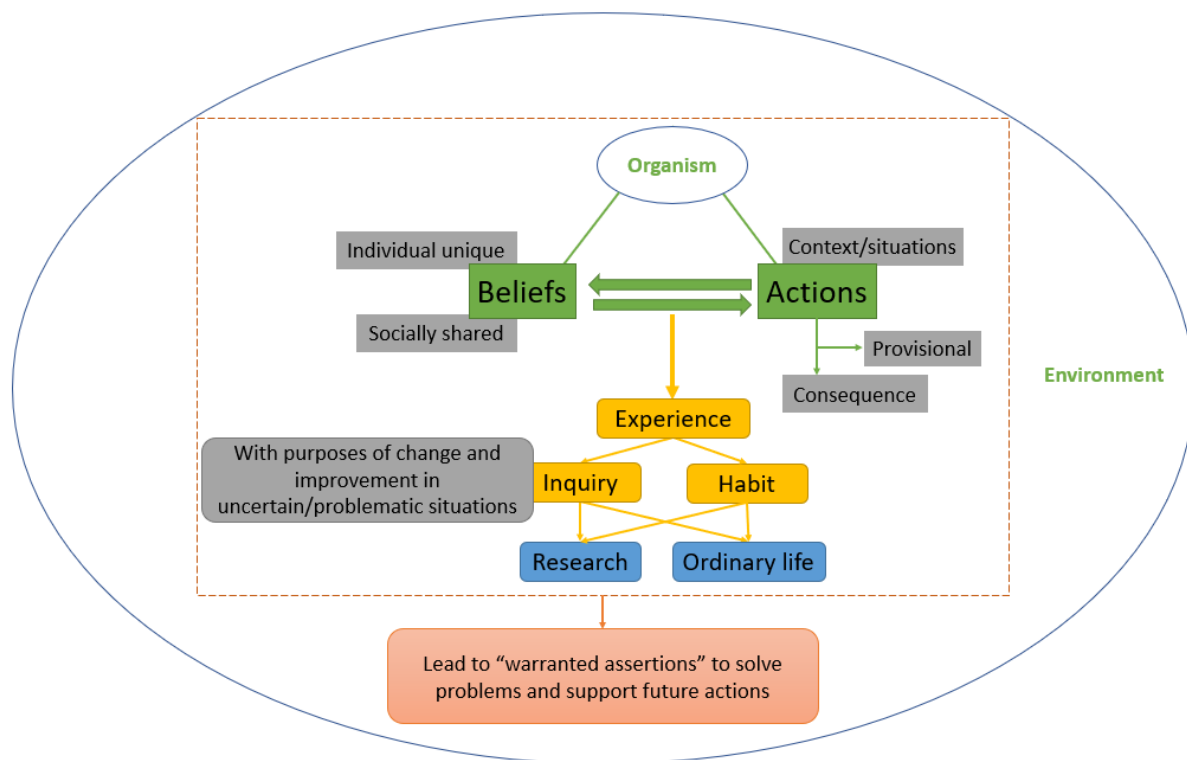


Figure 4.1 How Pragmatism understands the world

Pragmatists view the world as experiential of an existential reality, accept singular and multiple realities, and call for reflective research practice rather than merely mirror reality (Yvonne Feilzer, 2010). Knowledge is in the transaction between organism and its environment, which is both a construction and reality-based (Biesta & Burbules, 2003). The transaction between

**organism** and **environment** happens through a “doing-undergoing-doing” process, in which doing is based on previous **beliefs** and results in new beliefs that inform future **actions** (Dewey, 1939, p. 17; Biesta & Burbules, 2003; Morgan, 2014a). It is in this process, human being experiences. For pragmatism, beliefs are both individually unique that are based on one’s unique experiences and socially shared as it is generated from shared experiences (Kaushik & Walsh, 2019). Actions cannot be disassociated from the context they occur and thus the consequence of an action are provisional as the situation changes (Morgan, 2014b). Therefore, it is not that philosophy, or any theory tell people what and how to do, but real knowledge that fits with the current situation is derived through the process of action (Biesta & Burbules, 2003).

**Inquiry** as a form of experience is a core concept in pragmatism that happens in problematic situations and based on a series of self-conscious decision making, which is different from **habit** that happens semi-automatedly (Morgan, 2014a). Although problematic situation in ordinary life needs inquiry as well, **research** is a form of inquiry that requires more carefully and reflective decision making. Therefore, pragmatism is inherently a philosophy not only for research but also for ordinary life. It is by action and examining its practical consequences “warranted assertions” generated to help deal with the current problematic situation and decide which action to take next for future improvement, and the “warranted assertions” is provisional since context may change (Morgan, 2014a, p. 1048; Johnson & Onwuegbuzie, 2004).

As Dewey said,

“The ‘norms’ used at present have developed out of the processes by which metallic ores were formerly treated. . . Some procedures worked; some succeeded in reaching the end intended; others failed. The latter were dropped; the former were retained and extended” (Dewey 1938, p. 14).

Because of the focus on experience and the provisional nature of its resulted “warranted assertions”, pragmatists believe that pragmatism is not a recipe for educational research, and educational research does not provide prescriptions for what should be done in the future and what are the permanent truths in the field of education (Dewey, 1929). The philosophy of pragmatism and educational research only serve as instruments or resources that enable educational researchers or educators to have new insights toward the problems they might encounter and provide new possibilities of thought to inform their actions.

Pragmatism highly fits into and supports this study, and it exists, although implicitly, from the research purpose to data interpretation. Firstly, as mentioned in section 1.2, this study aimed to

explore the assessment of scientific argumentation competence (SAC) and to understand Chinese high school students' engagement in SA. So, this study takes SAC at the same time as measurable and generalizable and as participants' unique trait. Additionally, this study intended to understand Chinese high school students' engagement in SA by looking at their performance on an assessment and probing into their unique experiences. Secondly, SAC in this study was framed based on reviewing previous literature, which is consistent with the premise of the 'doing' in the transaction process, namely based on previous beliefs and aims to result in new beliefs that can inform the action in this study. Thirdly, as mentioned in section 3.5, this study is interested in both the process and the product of developing SAC assessment, and it is in the process that inquiry happens. Fourthly, SA itself corresponds to the philosophy of pragmatism, in which people construct, articulate, reflect on, and even modify their argument by transacting with others' opinions to reach the pragmatic goal of persuading or collaborating.

Fifthly, in order to assess SAC, this study tries to design an instrument and examine the consequence of using it. Section 4.2 will elaborate how the 'doing-undergoing-doing' process is embodied in the assessment development process in which both the researcher and the participants were experiencing by trying out, thinking, and reflecting. Sixthly, the focus of experience of pragmatism formulates Chapter 5 that does not aim to prove what is the truth or unchanging results, but to furnish some ideas and observations at a specific time and under specific context to help support future research. Lastly, the study does not aim to provide educational practitioners with direct ways to solve the problem but instead to provide them with information such as students' ways of thinking and their weakness in argumentation, although the final aim of the study is to help students improve their SA ability. Then teachers can have new perspectives towards SA and solve problems during the process of educational practice based on the information or knowledge generated in this study.

## **4.2 Research design**

Pragmatism, as the philosophical position of this research, commits to uncertainty by being open to shifts and changes in the relationships, structures and events being researched and being flexible to the use of a variety of types of data (Mounce, 1997; Yvonne Feilzer, 2010). From a pragmatist's viewpoint, what matters is whether the methods have the potential to answer what the researcher wants to know rather than excluding any particular methods (Yvonne Feilzer, 2010). This study seeks to offer a broad view in terms of SA and its

assessment by exploring a SAC assessment from both its process and product and exploring SA and Chinese high school students' SAC from both large-scale assessment and students' experiences. Thus, a mixed methods research approach, a research paradigm that acknowledges the value of multiple perspectives and ideas from both qualitative and quantitative research, is appropriate for this study (Johnson & Onwuegbuzie, 2007).

However, the mixed methods approach has been criticised for its requirement on the researcher to be skilled in several methods and for the 'whether and how can quantitative methods and qualitative methods be mixed' issue derived from its complex nature (Doyle et al., 2009). One of the core issues is how the different methods work together, which raises various typologies (e.g., fully integrated design, sequential design) for mixed methods approaches based on different criteria (Tashakkori & Teddlie, 2009). These criteria include but are not limited to the number of phases, priority of quantitative/qualitative methods, stage of integration of methods etc. (Teddlie & Tashakkori, 2006). But as Guest (2013) argued, these typologies fall short of providing a comprehensive perspective to support mixed methods research design for they don't capture the complex, fluid, and iterative nature of many mixed methods studies. One of the objectives of this study, namely developing an SAC assessment, determines that this study would be iterative and fluid, which makes none of these typologies for mixed method design precisely fit this study. Thus, inspired by Guest (2013), this study is concerned with justifying the research design by elaborating upon the *timing* and *purpose* of the integration of different methods, instead of giving a name to the research design.

Figure 4.2 shows the research design of this study. An initial version of a pencil and paper SAC test was designed and an iterative research procedure was thought to be useful to maximize the quality of the final assessment, as each iteration may generate different empirical evidence for how to improve the assessment and lead to different reflections upon it. Therefore, after each pilot study, a different version of the SAC assessment would be generated and be used as a research instrument in the next study. The iterative design was thought appropriate not only as it would make the process of developing the assessment a focus, but also as it would enable documentation of the process. A main study was included to help generate both large-scale and in-depth information, based on administering the test, to address the research aim. A **sequential** design (presented by green arrows) was required across/within studies because this could allow an iterative process for improving the SAC assessment, in which information obtained from the previous phase could **inform** the subsequent one. **Concurrent** design was considered useful

in that it could allow **comparison/complement** between the data from different approaches (blue box) within studies. The remainder of this section will clarify when and for what purpose the different approaches (i.e., qualitative or quantitative) were integrated according to each research question.

**Research Question 1**, which is '*How can a SAC assessment be designed for high school Physics students in China?*', aims at investigating the strategies that can be used to help develop SAC assessments by uncovering the problems and ways to address them in the process of modifying a pencil and paper SAC assessment. Small-scale interviews (i.e., teacher interview, student interview (follow-up), student think aloud) were considered helpful to provide in-depth and direct information (i.e., interacting with teacher/students directly) in terms of the potential problems with the assessment design. At the point when small-scale qualitative data could not provide useful information for improving the assessment quality, large-scale quantitative data would be useful by providing psychometric measurement characteristics of the assessment. So, within and across the pilot studies, the collection/analysis of each data set would **inform** the collection/analysis of the

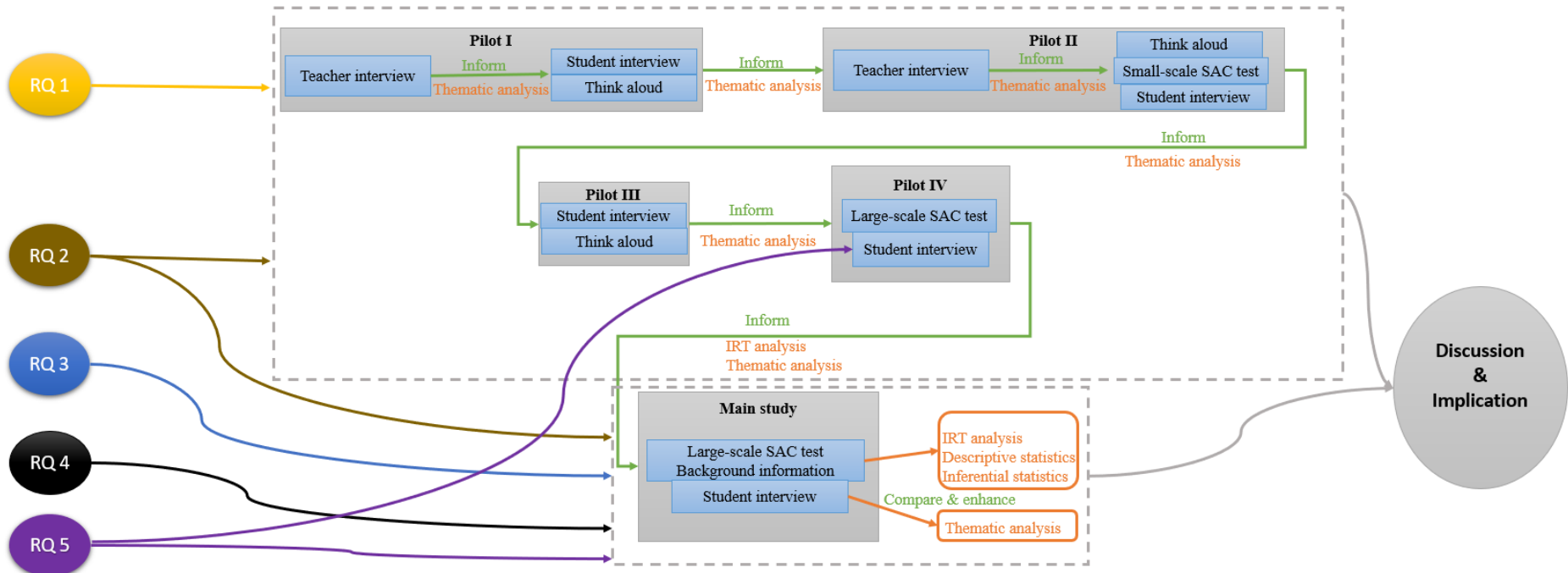


Figure 4.2 Mixed methods research design

next data set. The potential problems of assessment design revealed by qualitative data together with quantitative data and the strategies used to solve these problems would provide in-depth and broad perspectives to address RQ 1. The detailed assessment instrument development process and findings will be introduced in Chapter 5.

**Research Question 2**, which is *‘To what extent is the developed SAC assessment valid and reliable for assessing SAC?’*, is in nature the validation of the SAC assessment. As described in Chapter 3, validation can be taken itself as a study that embraces evidence from a variety of sources including test-takers’ voices. Additionally, both the process and product of developing an assessment are important. Therefore, both types of data in the pilot studies were considered to provide **complementary** evidence for validating the *process* of developing the SAC assessment. The large-scale test administration data and students’ follow-up interview data in the main study was thought to provide overall psychometric measurement characteristics of the assessment and in-depth information of students’ transactional experience with the assessment, these data could be **compared** with (or **explained** by) each other to validate the *product* of the SAC assessment. The resulted SAC performances of students after validation were thought to provide further insight of SAC as a learning progression and offer an overview of Chinese high school students’ SAC, which would answer **Research Question 3**, namely *‘What does the developed SAC assessment provide in terms of extended understanding of SA and of Chinese high school students’ SAC?’*. Chapter 6 will address RQ 2 and RQ 3.

**Research Question 4**, *‘How does the SAC of Chinese high school students as measured by the SAC assessment differ between different student groups?’*, was set to explore whether the potential factors chosen would influence students’ SAC performance to further understand their SAC. Quantitative data in terms of students’ SAC score, the school and class they were in, gender, presence of assessment scaffold (i.e., definition of SA provided on the SAC test papers), students’ Physics knowledge achievement and Chinese (i.e., language) achievement would be collected. Quantitative methods were thought appropriate to answer this research question because it would capture the possible relationships by inferring to the population, thus eliminating the possible bias of drawing a conclusion based on several participants. Chapter 7 will answer this research question.

To answer **Research Question 5**, *‘What are Chinese high school students’ perceptions of SA and the challenges they face in SA engagement?’*, qualitative interview data combined with the large-scale test administration data would be interpreted together. In this case, interviews

would be conducted after students took the assessment to explore participants' experience of taking the assessment and learning in schools, which could not be obtained from the test itself. Thus, qualitative data would not only be used to capture students' experience, but to understand the underlying reasons behind the scores that occurred. However, qualitative data have limited potential to make results generalizable. Results from quantitative data would provide information from a broader perspective in terms of the challenges faced by the population. Therefore, both types of approach could be **combined** to offer broader perspective to address this research question. Chapter 8 and part of the results in Chapter 6 will help uncover this research question.

Overall, qualitative data and quantitative data were designed to complement, corroborate, and enrich each other to help achieve the research aim (Johnson & Onwuegbuzie, 2007).

### **4.3 Sampling**

This study adopted a mixed methods sampling approach because of the need to obtain data that could be both generalizable (quantitative) and in-depth (qualitative) and the data structure was nested to reflect the education system (i.e., students within classes, classes within schools etc.) (Teddlie & Yu, 2007). So, the sampling approaches for cities, schools, classes, and students were different. Despite including large-scale data intended to represent the overall situation of Chinese high school students, the limited resources and time determined that this doctoral study could only get close to this goal.

#### **4.3.1 Overall consideration**

In general, the overall education level in economically developed provinces is better than that in less developed provinces in China due to the influence of economy and the resources it brings. Overall, provinces in the east of China are more developed than provinces in the west of China; and provinces in the south of China are wealthier than those in the north of China. Big gaps exist between urban areas and rural areas (OECD, 2020). However, all the provinces use the same Curriculum, which makes it feasible and reasonable to invite students across provinces to take the same assessment since students are supposed to learn similar content. Due to the limited accessibility of contacts in the west of China, this research only considered provinces in eastern China. Provinces in both the north and south of eastern China were selected so as to generate data that are more representative of the situation in China as a whole. Figure 4.3 below shows the location of the sample in this research.





Figure 4.3 Locations for data collection in China

Overall, homogeneous sampling (a form of purposive sampling) was used for the selection of grade 11 (aged around 16-17) students across each stage of the study to reduce variation across grade levels (Etikan et al., 2016). Grade 11 was recruited because it was more feasible given students in grade 12 were preparing for the college entrance examination (Gaokao). And grade 11 students have learned most of the modules in high school compared with grade 10 students who have just enrolled in high school.

#### 4.3.2 Sampling for Pilot I to Pilot III

A convenience sampling approach was used to invite students in the first three pilot studies because only a small number of participants were required for each study to obtain in-depth information, and to save time and resources. Students in the first and the second pilot were from public high schools, and students in the third pilot were from a tutoring institution (i.e., New Oriental Education & Technology Group). These students were recruited for the accessibility of gatekeepers (i.e., their teachers) and students' willingness to participate. Both convenience and purposive sampling approaches were used to invite teachers (Etikan et al., 2016), because this research aimed to invite novice and expert physics teachers in high school and science education researchers to help review the assessment instrument. The reason for

inviting teachers in different positions was to alleviate the limitation that there were few SA experts in China and limited accessibility to SA experts. The sample size of each pilot study is illustrated in Table 4.1 below.

*Table 4.1 Sample of the first three pilots*

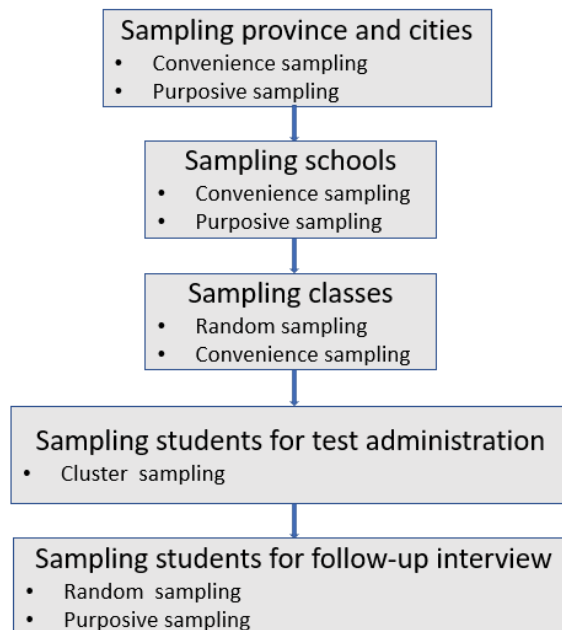
<b>Pilot</b>	<b>Teacher</b>	<b>Student</b> (girls/boys)	<b>Province /City</b>
<b>1</b>	1	4 (2/2)	Jilin/Changchun
<b>2</b>	10	30 (Think aloud: 2/2 Follow-up: 2/2)	Jilin/Changchun
<b>3</b>		11 (7/4)	Beijing

Pilot I aimed to try out a few item examples, so only 4 students were invited at first. The information obtained from these four students was enough to help modify the assessment, so no more students were invited at this stage. To avoid the possible influence of gender on students' performance on the items, two boys and two girls were invited. Additionally, the teacher was asked to invite one boy and one girl each with an above-average school Physics achievement and one boy and one girl with a below-average school Physics achievement to mitigate the possible influence of school achievement on their engagement in the items.

Pilot II aimed to obtain feedback of a complete test, so in addition to think aloud and follow-up interview, 30 students were invited to check whether there are items that are too easy or too difficult for the students. Pilot III aimed to prepare the test for a large-scale administration, so 8 students were invited to think aloud to make sure that the test was appropriately designed, and 3 students were invited to take the test to check the time they needed to finish the test. Since the first two pilot studies found no gender differences, the numbers of boys and girls invited in the third pilot were not strictly balanced.

### **4.3.3 Sampling for Pilot IV**

For the fourth large-scale pilot, Figure 4.4 below shows how a mixed methods sampling approach was employed.



*Figure 4.4 Sampling approach for the fourth pilot*

When this pilot study was conducted, the pandemic made traveling a complex issue, so Shandong province was chosen because it's my home province and the gatekeepers were more accessible given the collection of large-scale data. In China, schools in bigger cities are generally of higher quality than schools in smaller cities because they have more resources in terms of facilities and teachers. To make the sample as representative of the province as possible, cities were selected purposively to include the capital city (i.e., Jinan), an ordinary city (i.e., Jining), and county-level city (i.e., Qufu). In China, the performance level of a high school is often decided by the percentage of students that enrol in university in each year. High schools in a city/province are usually compared by the approximate average of these percentage values. Therefore, there was not a precise ranking of the schools within a province/city, but the relatively position of a school (i.e., top, upper-middle, below-middle) is often stable over a several year period. Information on school performance levels in this study was obtained from teachers at each school. One of the two invited schools in Jinan was one of the top high schools in Shandong province, and the other one was an upper-middle high school in Jinan city. The schools in Jining and Qufu were both top high schools in their respective cities. Due to the limited access to lower performing schools, the sample may result in higher outcomes than the average student performance in the province.

As for participants, in order to save time and resources, cluster sampling was used to choose

two or more classes in each school (Teddlie & Yu, 2007). For schools in which the school principals were contacted and agreed to take part, classes were invited randomly; while for schools in which teachers were contacted, the classes that were taught by the specific teacher were invited. Students in all invited classes were invited to take the exam and their right to withdraw from participation were reserved. After collecting test papers, students' test papers were checked by looking through their responses. However, the test papers were not scored at that time because scoring takes time whereas participants should be invited for interview within a short time of finishing the test so that they could have a reliable memory of taking the test. Within students who were willing to participate in the follow-up interview, those who provided test responses of three general types (i.e., responses with blanks or scribbled words, rich and logical responses, and responses between the two extreme ends) were randomly invited to the interview. There were in total 373 test papers collected, and 18 students were invited to the interview. A description of the sample for the fourth pilot is shown in Table 4.2 below.

*Table 4.2 Sample of the fourth pilot*

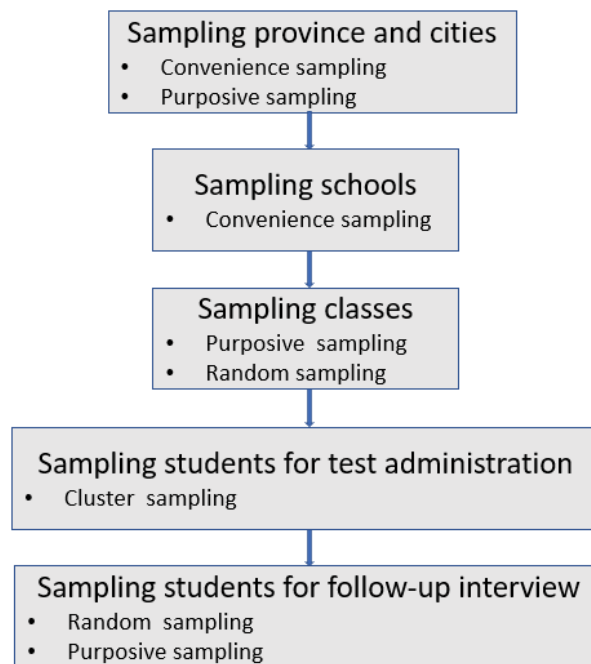
<b>City</b>	<b>School</b>	<b>Class</b>	<b>Students</b>	<b>Gender (girls/boys)</b>	<b>Follow-up interview (girls/boys)</b>
Jinan	A	2	85	24/61	6 (3/3)
Jinan	B	2	44	7/32	6 (2/4)
Jining	C	3	90	19/61	4 (1/3)
Qufu	D	4	154	62/92	2 (0/2)

Participants who were finally interviewed in this pilot were based on their willingness to participate, their time schedule, their access to a mobile or a computer, and my time schedule. So, the number of students participated in the interviews were not the same across schools.

#### **4.3.4 Sampling for the main study**

In terms of the main study, Figure 4.5 below demonstrates the sampling approach. More students were invited to take the test because the main study did not only aim at obtaining the psychometric information of the assessment but to depict an overall picture of Chinese high school students. The selection of cities was based again on a convenience and purposive sampling approach. For Jilin province, cities were selected purposively that included the capital city (Changchun), ordinary cities (Shulan and Jiaohe), and a county town (Nongan) to give a wide cross-section of schools. For Guangdong province, only Shenzhen city was selected due to the accessibility of schools in that city. As with the fourth pilot, the schools in each city were either top high schools or upper-middle high schools in that city due to limited access to

low performing schools.



*Figure 4.5 Sampling approach for the main study*

It is quite common in China that students are assigned into different classes based on their achievement performance in the entrance examination, namely, the *key* classes and the *ordinary* classes. Key classes are better classes in a school, and the students in the key classes have better achievement on their entrance examination and the university admission rates of the key classes are generally much higher. Given that the main study required a large amount of data and aimed to provide an overview of SAC among Chinese high school students, for schools that agreed, key classes and ordinary classes were deliberately included and randomly invited. Otherwise, classes were randomly invited within ordinary classes (schools didn't like students in key classes having to spare time on activities not related to Gaokao). Depending on the number of classes in each school, one to four classes were chosen randomly to provide two types of test papers (i.e., with assessment scaffold and without assessment scaffold) to students within each class randomly. The sampling scheme of students within classes and for interviews were the same as that in the fourth pilot.

In total, 1413 test papers were collected in the main study and 12 students were invited to the follow-up interview. Table 4.3 shows the sample size in the main study.

Table 4.3 Sample of the main study

Province	City	School (identifier)	Class (key/ordinary)	Students (number)	Gender (girls/boys)	Follow-up interview (girls/boys)
Jilin	Changchun	5	3(1/2)	128	64/62	2 (1/1)
	Shulan	6	4	195	76/117	1 (0/1)
	Jiaohe	7	2	52	20/32	
	Nongan	4	2(1/1)	119	57/56	3 (1/2)
		3	11(5/6)	401	223/160	
Guangdong	Shenzhen	2	5(2/3)	133	60/70	2(1/1)
		1	11(3/8)	385	165/204	4(2/2)

Overall, although the large-scale sampling, especially that in the main study, tried to make the sample diverse in terms of the area in which schools were located, levels of cities, and levels of schools and classes, the sample was still limited in its representativeness because it did not include schools from less developed areas or schools with poor performance within a city. Additionally, the sample sizes in the fourth pilot (i.e., 373) and the main study (i.e., 1413) are based on the valid test papers, not the complete set of test papers that were collected back (i.e., 400 in the fourth pilot and 1668 in the main study). This is because test papers that were blank (i.e., without any information or with obviously irrelevant information), with 7% in the large pilot and 15% in the main study, were excluded from the data set in the first place. So, another limitation is that the outcomes based on the sample might be higher than for the population in general as a result.

#### 4.4 Data collection

This section provides information of the data collection strategy used in this study and outlines the procedures employed for data collection.

##### 4.4.1 Data collection strategy

This section critically justifies the data collection strategies used in this research by considering the benefits and caveats of using each strategy. These strategies explored below include tests, think aloud and semi-structured (follow-up) interview.

###### 4.4.1.1 Test

A test, as a data collection strategy, is a variety of techniques designed to assess knowledge, ability, or intelligence (Tashakkori et al., 2020). Tests can be used to collect both qualitative data and quantitative data. Close-ended items like multiple choice items produce quantitative

data (i.e., TEST-QUAN data), while open-ended items like essays generate qualitative data (i.e., TEST-QUAL data). Usually TEST-QUAL data are transformed to quantitative data (marks awarded) when the research needs numeric information (Tashakkori et al., 2020).

Rather than using an existing test to collect data, this study aimed to design a test based on the understanding of SA in this study. As shown previously in section 4.2, several pilot studies were conducted, and a test was not only used for large-scale administration but also as a tool to elicit interview data on what may contribute to a better design of a SA assessment. That is, the tests used were modified iteratively during the research process and different version was used in each phase of the research. The detailed procedure of how a test was developed and the information within the test will be introduced in Chapter 5.

In addition to the demanding effort of developing a test, preparing examinees and test administration are key phases in collecting testing data (Lane et al., 2015). The appropriateness of test preparation activities depends to a large extent on the context in which the test is designed and implemented (Mehrens & Kaminski, 1989). Extensive test preparation may complicate the interpretation of test results thus bring worse effect to test validity (Bishop & Davis-Becker, 2015). The test in this study was not high-stakes test and didn't intent to include any content knowledge that goes beyond what the students have learnt, and the item types were designed to be familiar to the students. So, as previous studies that aimed to assess SA (Osborne et al., 2016; Lee et al., 2014), this study didn't plan to train the test takers before asking them to take the test but focused on making sure that students know the purpose of the test (Bishop & Davis-Becker, 2015).

Except for random errors, tests should be carefully administered to minimize the construct-irrelevant variance thus to reduce systematic errors (McCallin, 2015). Sources of construct-irrelevant variance when administering a test include Physical environment, Instructions, equipment and support, Connectivity, Time limits and speededness, Test administrator effects, and Fraud and security (McCallin, 2015). The test format (i.e., pencil and paper test) and the location of administering the test (i.e., classrooms that the students are usually in when taking school tests) were thought to help avoid construct-irrelevant variance caused by testing environment. Teachers were asked to monitor in the classrooms to avoid fraud. Time limits needed to be decided by balancing the need of including enough items to represent the construct of SAC, the feasibility of administering a test in schools, and the time students need to finish such a test at their age/grade. Thus, after the iterative piloting, the test length was finally

adjusted to that student can finish at around 45 mins, which is the duration of one class. In addition, the researcher was not allowed to enter the classrooms during COVID-19 pandemic and thus teachers were asked to help administer the test, additional information was considered helpful to check the administration of the test. So, a short survey was included in the test as shown below:

- 1) Was the time sufficient to finish the test? (yes/no)
- 2) Did you take it seriously when doing the test? (Very serious/serious/not sure/not serious/very not serious)
- 3) Did you finish the test independently? (independent/ partly independent/ not independent)
- 4) Do you think the test is difficult? (Very difficult/difficult/not sure/not difficult/very easy).

Background information such as gender and age, and whether students were willing to participate in a follow up interview (leave their contact number if so) were also collected with the test papers.

#### **4.4.1.2 *Think aloud***

Think aloud, also called cognitive labs, is a data collection strategy that investigates participants' thinking processes and has been considered one of the most effective ways to assess higher order thinking processes (Wilson, 2004; Olson et al., 1984). During the think aloud process, participants think out aloud when performing a task. In test development, it can provide detailed and timely information of the response process of test takers. Although similar information can be obtained from a follow-up interview that is conducted after participants take the test, the delay might interfere with the respondent's memory and thus provide less detailed information of the response process (Wilson, 2004). Therefore, think aloud was used during the test development process to collect data to inform the modification and improvement of the test.

Nevertheless, think aloud is not omnipotent. According to Vygotsky's (1962) theory about inner speech and abstract thought, the parts of thought that can be translated into language are only part of the complex thought network. Even think aloud, which translates thoughts into words, cannot uncover deeper thought processes in their true complexity (Charters, 2003). Thus, a think aloud strategy needs to be used carefully. Specifically, verbal reports should immediately follow the thought process to provide accurate information; it is better to ask participants to think aloud during a specific task rather than in a general way (Eccles & Aarsal,



2017); either thoughts that happened naturally and effortlessly, or that happened deeply and abstractly might not be verbalized (Charters, 2003). Thereby, tasks that are highly cognitive demanding or extremely simple are not suitable for think aloud (Ericsson & Simon, 1980). Researchers recommend that tasks that require “cognitively demanding language use” and are separated into task units where one unit is worked at a time are appropriate for using think aloud strategies (Charters, 2003, cited from Akyel and Kamisli, 1996, p. 15-16).

In order to reduce the potential problems of using thinking aloud and maximize the useful information it provides this study took the subsequent considerations into account. In this study because the participants had no idea of what a think aloud method is and how to think aloud, the researcher explained to them about think aloud and demonstrated to them how to think aloud using a real-life scene related example before asking them to think aloud themselves. As asking participants probing questions during the think aloud process can distort their real thoughts (Sugirin, 1999), they were told that they would not be interrupted or responded to during the process. Before they thought aloud, they were asked to speak about everything in their mind including their feelings, and the researcher would encourage them to keep thinking aloud by saying “keep talking” when there was a long period of silence. Additionally, participants were asked to notify me after completing each task’s thinking aloud. Then the researcher asked participants retrospective questions about their think aloud of a task to:

- 1) elaborate ambiguous utterance and validate my understanding of their expression,
- 2) add depth to information,
- 3) obtain information from participants who had difficulties in thinking aloud, and
- 4) familiarize participants with think aloud and have a break before proceeding to the next task.

Before going to the next task, participants were asked whether they felt comfortable about thinking aloud and whether they would like to continue. Each participant in the think aloud thought aloud about no more than 3 tasks because of the cost of time and energy.

#### ***4.4.1.3 Semi-structured interview***

Interview is widely used in mixed methods research since it can generate both qualitative data (i.e., by using open-ended questions) and quantitative data (i.e., by using close-ended questions) and is viewed as a powerful data collection strategy as it entails one-to-one interaction between researcher and participant (Tashakkori & Teddlie, 2009). Open-ended questions in interviews

can produce in-depth data that enrich or complement quantitative data. Semi-structured interviews have some predetermined questions while allowing for flexibility in how the topic is addressed (Dunn, 2005, cited from Longhurst, 2003). Therefore, the approach cares about participant's experience of a situation about which there is sufficient objective knowledge but less subjective experience (McIntosh & Morse, 2015). Semi-structured interview was used in this study because there have been plenty of research discussing SA theoretically, and there have been studies conducted in China analysing students' SA, but there is a scarce of research concerning students' experience and perspectives of engaging in SA.

In this study, follow up interviews were conducted after students took the test, taking the form of semi-structured interviews with open-ended questions (Qu & Dumay, 2011). Semi-structured follow up interviews served as an effective tool to understand how students felt during the assessment and what the experience was like, how they perceived SA, how they experienced their school learning, and how this affected their SAC. This information would not be obtained directly from the test. The follow up interviews aimed to collect two types of data:

- 1) data that relates to how participants got their answer on certain items to examine the items, aiming at addressing RQ 1 and RQ 2, and
- 2) data that involves their experience of learning, argumentation and taking the test, aiming at addressing RQ 5.

Interviews conducted in the first three pilot studies were focused on obtaining the first type of data, complemented with think aloud interviews to get more detailed and accurate information. In contrast interviews conducted in the large-scale pilot and the main study obtained both types of data to enrich the understanding of students' SAC and their related experience. All the interviews were audio recorded.

Despite the semi-structured nature of the interview, not all the students were asked exactly the same questions. This was because participants who showed difficulty in recalling/elaborating some questions or who were excited in talking about their own experiences were not pushed or interrupted, and some questions were asked based on their responses to the test items. In the first three pilot studies, participants were invited to the follow-up interview within 1 day of finishing the test, while in the large-scale pilot and the main study, students took part in the interview within 7 days of taking the test. Participants who accepted the invitation to the

interview with more than 7 days delay or who were unable to talk (two of the invited participants preferred to text message) were politely declined to participate in the interview. This was to ensure that their recall of the experience was accurate and in detail. The pre-outlined interview questions are listed in Appendix 3.

#### **4.4.2 Data collection procedure**

Because of Covid-19, I did not get a chance to go back to China for the first two pilots, and I was in self-isolation in China when conducting the third pilot. During the fourth large-scale pilot and the main study, I was able to enter the schools and talk to the teachers or school principals face to face but was not allowed to go into the classrooms. Therefore, all the test papers were administered by teachers in the school and all the interviews were conducted online. Another thing to note is that all the students chose to participate in the interview via voice call, rather than video call. All the interviews lasted around 35 to 60 minutes. Given that several studies were conducted iteratively in this research, this section discusses the procedures of data collection by introducing each study separately.

##### ***4.4.2.1 First pilot***

The first pilot was conducted in March 2020. One high school physics teacher was provided with the assessment framework of SAC and several initial test items to review a week before the interview. During the interview, the teacher provided suggestions and discussed with the researcher the assessment design. The test items were modified based on the teacher's feedback. After that, 2 students were invited to the think aloud and another 2 students participated in the follow up interview after they finished the test. These students were in the same class in the same high school, and they were taught by the Physics teacher who participated in the previous interview. During the interview, I spent several minutes to chat with them to make sure that they were not facing any tension. The pilot had a small number of participants, the reason for this was to save time and resources, and information provided by these participants was already useful by itself to support modifying the assessment at the initial stage. Based on the results of this pilot, a complete version of the test was designed.

##### ***4.4.2.2 Second pilot***

The second pilot was carried out between May 2020 to July 2020, both teachers and students were invited in this phase. More participants were invited to this pilot, as a complete version

of the assessment was now designed and the quality of the revised assessment should have improved, so it was thought that it would be more difficult for only a small number of participants to provide useful information. Taking the form of an *item panel* in Wilson's (2004) four building blocks model, based on their willingness and time schedule, 10 teachers were invited to review the assessment instrument either individually or joining in a panel including several other teachers. As in the first pilot, teachers were provided with all the related information including the background of the research, the assessment framework, test, and scoring rubrics at least one week before the interview. The tasks in the test were discussed one by one during the review meeting. As in the first pilot, the test was modified according to the teachers' review.

After that, four students were invited to think aloud, and another four students were invited to the follow up interview after completing the test. To prevent students from taking too much time to finish the think aloud, students responded to different subsets of items to make sure that all the items were thought aloud by around 4 students. A small-scale test administration was needed to obtain information such as the frequencies of students' responses to test items (to see if there were items that were extremely easy or difficult), and the time needed to finish the test. Therefore, 30 students excluding the 8 students who took part in the interview were administered with the test paper by their head teacher. These 8 students were asked not to share the test items with their classmates. Then, the teacher scanned these test papers and sent the electronic copies to me. All the students in this phase were in the same high school and taught by the same Physics teacher. By analysing all the data collected in this phase, the test and its related materials were revised.

#### **4.4.2.3 *Third pilot***

The third pilot was conducted during September 2020, in which more students were recruited for thinking aloud to obtain more information in terms of how students respond to the items and to prepare the assessment for a large-scale administration. Specifically, 8 students joined in the think aloud and 3 students were invited to complete the test. However, only 2 students after taking the test participated in the follow up interview and the other student withdrew from the interview. The participants in the third pilot showed sufficient engagement with the items and understood the items in the way they were intended. So, after modifying some language issues, the test was prepared for the fourth large scale pilot.

#### **4.4.2.4 Fourth pilot**

The fourth pilot happened in October 2020. After printing test papers, I firstly went to two high schools in Jining city and talked with their principal about the research. Both schools were willing to participate in the study, but only one of these two schools was finally contacted since few students in the other school learned Physics. Due to the restrictions to enter the classrooms at the time of this pilot, test papers, information sheet, and consent form were handed to the principal of that school and the principal was asked to recruit classes randomly. This pilot didn't aim to do statistical analysis for students' SAC score and other variables such as whether they took a test with assessment scaffold, because this pilot might not obtain valid test scores since items might need to be modified. And providing students with assessment scaffold (i.e., definition of SA elements) could help reduce students' unfamiliarity with SA. So, each school was given only about 30 test papers that do not have scaffold to obtain interview data in terms of how scaffold may help students better understand SA. Like their normal school test, students were given 90 minutes to take the test in their classrooms. Based on the principle of volunteerism, they were also informed by teachers that if they were not willing to take the test, they could do their homework without interfering with other students. The time of testing was decided by the principal according to the school's schedule. After the test was over and the test papers were submitted, the principal contacted me to pick up the test papers.

As elaborated in section 4.3 and 4.4.1.3, within 7 days of collecting these test papers, students who provided different levels of responses were invited to follow-up interviews. The procedure of collecting data in other schools were similar. The principal of the school in Qufu was contacted and helped me administer the test paper by recruiting classes randomly. For the two schools in Jinan, teachers were contacted so that it was the classes they specifically taught that participated in the study.

#### **4.4.2.5 Main study**

The main study was carried out during December 2020. The data in this phase were collected from two provinces which were far away from each other, and time for collecting data was limited because schools were about to start preparing for end-of-term examination before the Chinese New Year. So, I only went to Shenzhen city myself, whereas the data collection in Jilin province was entrusted to teachers in each school. For the data collection in Shenzhen city, test papers were printed, divided based on the approximate number of students in each class,

and handed to the principals of the two participated schools. The detailed procedure was similar to that in the fourth pilot.

In terms of the data collection in Jilin province. A colleague who was based in Jilin province helped me printed all the materials and posted them to the teachers in each participating school because I was collecting data in Shenzhen city at that time. Teachers in each school posted the test papers back to the colleague after students finished it (Figure 4.6). The colleague randomly selected some test papers and scanned them to me so that I could invite students to follow-up interview. This procedure was adopted because it takes less time to send a courier within the province, so students could be invited for a follow-up interview within 7 days of completing the test.



*Figure 4.6 Test papers posted to and collected back from schools in Jilin*

All the schools were provided with the same material and were informed about the same guideline about how to administer the test. As previously mentioned, schools' schedule at that time was tight because of the upcoming New Year holiday and end-of-term examination, most schools contacted agreed only one lesson's time (45 minutes) for participating the research. In addition, almost all participants in the fourth pilot used less than 60 minutes to finish the test and the test in the main study includes less items. Therefore, 45 minutes was appropriate for students to finish the test and was convenient for schools. The main study also aimed to explore the possible influence of providing students with SA definition (i.e., assessment scaffold) on their SAC performance. So, the guidelines in the main study were:

- 1) Please make sure that students do not communicate with each other during the test.
- 2) The time for finishing the test should be 45mins.

3) Please randomly select the classes that will be administered with both types of test papers (with/without assessment scaffold), and please distribute both types of test papers randomly to the students within the selected classes.

It was intended to provide both types of test papers to a subset of classes and providing the test paper with scaffold to all other classes to reduce the possible threat of the students' unfamiliarity with SA to validity. The reason to distribute both types of test papers within a class was to compare the students' (who are in the same class) performance on the two test types to eliminate the possible variance between classes. Nevertheless, some teachers didn't operate correctly, and they didn't assign the two types of test papers within a class rather assigning either type to a whole class (see section 7.3). Other guidelines in terms of the voluntary nature of the research will be discussed in section 4.6.

Considering the close relationship between content knowledge and SA, and SA being mediated by language (as discussed in Chapter 3), RQ 4 also considers how Physics content knowledge and Chinese performance may influence students' SAC performance. However, designing another test for Physics and Chinese is a massive task that need extra expertise on both field, which goes away from the main aim of this research. So, teachers were invited to supply scores from the latest assessment of Physics and Chinese which may be used as indicators of Physics knowledge and ability to understand and write. Considering students' achievement was developing, the most recent test records was thought to represent the students' achievement level at the time of taking the SAC test more precisely. However, only three schools provided any of this information, and only one school provided the records of the Chinese test.

## **4.5 Data analysis**

This section shows how the data were analysed by explaining the analysis of data collected by think aloud, semi-structured interview, test paper, and school achievement test respectively. Although thematic analysis was applied in both think aloud and semi-structured interviews, the focus and procedure of analysis was different. So, the analysis of data collected from these two strategies will be introduced separately.

### **4.5.1 Think aloud data analysis**

The analysis of think aloud data was considered highly pertinent to the assessment development, thus was used to help answer RQ 1 and RQ 2. As mentioned in section 4.4.1.2,

the think aloud method is concerned with the thinking process of participants and many studies have used it to explore patterns of thinking in problem solving (Eccles & Aarsal, 2017). However, given the aim of employing the think aloud method in this study was to detect any problems in the test to improve it, this study didn't pay attention to the various thinking patterns participants might demonstrate during think aloud. Rather, from the beginning of the interview when analysis had started implicitly, the focus was on whether students are displaying argumentation and whether there is any problem with the test. Like Kvale (1996) mentioned the conversation between researcher and participant starts in the interview situation and continues to the analysis of the transcripts. Specifically, the researcher brought several questions in mind to the think aloud data collection and analysis process:

- 1) how they got their answer
- 2) why they got the item right or wrong
- 3) whether and how the design of the test/item affected their performance.

Although the researcher aims to answer the three questions, the generation of codes were induced from the data. Specifically, this study adopted Braun and Clark's (2006) six-phase approach to thematic analysis. These six phases are

- 1) familiarizing yourself with the data
- 2) generating initial codes
- 3) searching for themes
- 4) reviewing potential themes
- 5) defining and naming themes
- 6) producing the report.

As Braun and Clarke (2006) mention, thematic analysis may start during the data collection. This is especially the case for the think aloud data given think aloud data is more difficult to understand because it is not usually expressed in complete and reasoned sentences like in writing and speech (Charters, 2003). Therefore, during the data collection process, without interacting with the participants, what they said were carefully listened to and compared with the item and notes were taken to identify what was relevant to the research question. As mentioned previously, participants thought aloud task by task, and in between, they were asked to review what they said for confirmation. After finishing each interview, the test design was reflected on combining with the notes during the interview to contribute to test modification.



Data were transcribed and transcriptions were checked by listening to the recording again.

As mentioned in Chapter 3, there is a scarcity of research exploring how to design SA assessments, so a more inductive approach was adopted when identifying important information in the data (Braun & Clarke, 2006). However, the data analysis during the think aloud process was on the specific aim of finding out the possible problems with the test, this specific aim provided directions for data analysis. Additionally, some widely discussed problems of designing assessments such as language and content knowledge were already being considered when analysing the data.

In terms of the second phase, the transcripts in the first pilot were manually coded because only two students were interviewed, while the Nvivo software was used for coding in the second and third pilot. Given the procedure of thinking aloud was task by task, an identifier such as 'Task 1' was created for each task to include the codes generated by thinking aloud the task. Other information that did not relate to specific tasks was coded separately (please see the screenshot of Nvivo in Appendix 4). To obtain as much information as possible, any information that might be relevant to the research question was coded when generating the initial codes. The low frequency for codes shown in Appendix 4 is due to there were only 4 students participated in think aloud in the second pilot and the same code may be generated several times but located under different task identifiers. For example, the theme 'Understanding about the problem being argued' was constructed by combining codes 'misunderstand the test aim', 'misunderstand the topic being argued', 'without paying attention to the problem being argued', and 'confused about the topic being argued'. Based on the theme and codes, the strategy used to solve this problem is 'making the problem to be argued explicit' (see section 5.4.2). The findings from analysing think aloud data and the changes it brought to the test design will be demonstrated in Chapter 5.

As for generating and reviewing themes, rather than expounding a rich story, the analysis of this phase was focused on summarizing the factors that undermined the test quality to inform the test modification in this and future studies. Thus, only codes that related to the aim of the analysis were considered to construct themes. It was easy to recognize some apparent problems residing in each task, like language, but to find deeper problems, it was necessary to relate all the codes from different tasks. The renaming of themes and writing went hand in hand until all the factors were considered and elaborated upon.

#### 4.5.2 Semi-structured interview data analysis

As mentioned previously, two types of data were collected in the semi-structured follow up interview:

- 1) focusing on how students figured out certain items and
- 2) exploring their experience related to SA, science learning and test taking.

The follow up interviews in the pilots mainly collected the first type of data and aimed to improve the test quality and so the data were analysed together with the think aloud data adopting the same procedure and with the same focus. The fourth pilot and the main study collected both types of data with the main study focused more on the second type of data. Thus, this section demonstrates the analysis of part of the interview in the fourth pilot (the second type of data) and the interview in the main study for the aim of understanding students' SAC, which was mainly to be used to answer RQ 5.

The analysis for the semi-structured interview again adopted the six-phase approach of Braun and Clarke (2006). Unlike the analysis of the first type of data in the pilots, which focused on identifying the problems of the test, the analysis in the main study aimed to explore students' experiences of SA engagement and learning science. So, from the data collection to the data analysis, instead of having a specific and clear aim as in the think aloud data, a more open mind was provided to communicate with participants and analyse the data. Right after the interview with each participant, my experience of talking to that participant such as the impressions of the participant and interesting points found from the interview were documented. In addition, all the interviews were transcribed, and transcriptions were checked by listening to the audio again.

Nvivo software was used for analyzing data. Due to there being more participants in the fourth pilot (N = 18) and in the main study (N = 12), the transcriptions of 2 to 3 participants were coded and reviewed at first before proceeding to the remaining transcriptions. This procedure was helpful since it familiarized the researcher with coding the data and made the coding for each subsequent transcription occur in a more systematic way. Annotation was made during the coding process to record ideas coming out from the coded data that may contribute to themes or to the research aim. After coding each transcription, a corresponding memo was made to summarize any overall characteristics worth noting of each participant or any thoughts inspired by coding that transcription. The annotation and memos were thought helpful in

generating themes or discussing storylines for Chinese high school students' SAC.

Unlike for the data analysis in the think aloud, themes were constructed in this phase to generate a rich description that elaborates Chinese high school students' SAC. Possible themes and the relationships between them were drafted by reading codes and transcripts. The initial writing started after drafting a theme map and formulating the story in mind. During writing, codes and transcripts were re-read and themes were reviewed, which reformulated the writing. The construction, review, and renaming of themes and writing/rewriting reports were conducted recursively and interchangeably (see Appendix 5 that shows the drafts/notes when constructing themes at different stage). The process continued until the writing was coherent and answered the research question precisely. Appendix 6 demonstrates a screenshot of the codes in Nvivo (the codes were translated into English during the writing process), which are from the main study.

Another thing to mention is that the language used from transcripts to codes and drafted themes was all Chinese. This was to avoid possible distortion of meaning during interpreting data that were collected in Chinese. Before starting to write, the codes and themes were translated to English and results were written using English. Because I have been very familiar with the data and the patterns in the data until this stage, the recursive revision of writing and themes were using Chinese and English interchangeably. The translation of the final themes and corresponding illustrative codes are shown in Appendix 7.

### **4.5.3 Test data analysis**

The analysis of SAC test scores was used to revise and validate the test and provide further insight into the structure of SAC, thus was used to help answer RQ 1, RQ 2 and RQ 3.

#### **4.5.3.1 Scoring**

Responses for the open-ended items were rated by two raters based on the scoring rubrics (see section 5.3.3 for the design of scoring rubrics). Except for the researcher, the other rater was a colleague who was a teacher at a university in China and was familiar with this research. The same two raters scored the test papers in the two large-scale studies (i.e., fourth pilot and main study). Since mainly E-SA items were modified after the fourth pilot, the open-ended items and their scoring rubrics were almost the same in the two studies. Raters adopted an iterative procedure of scoring in the fourth pilot. We firstly scored a small part of the test paper and then

discussed and reached agreement on the specific scoring standard of each item. At this stage, 11 test papers were scored together. After an agreement was reached, we marked 94 test papers together and resolved any disagreements. As for the remaining test papers, the researcher scored them alone to save resources (Lee et al., 2014). To ensure consistency, the remaining test papers were scored twice. There were 98 test papers scored by both raters in the main study.

#### **4.5.3.2 Item response theory analysis**

After obtaining the scores of each participant on each item, Item response theory (IRT) analysis was employed to show statistical characteristics of the test and each item. IRT is also called latent trait theory or modern test theory and its analysis unit is the item rather than a sum score, which is different from True score theory (TST) or Classical test theory (CTT) (Paek & Cole, 2019). There are several IRT models that share common features that advantage IRT over CTT, for example, it provides information for each item and estimates item parameters/person ability independent of the specific sample of respondents/items (Bond & Fox, 2015; Paek & Cole, 2019).

The fourth pilot aimed to explore any potential problems in the items by describing the feature of the data. Factor analysis was used to check whether the data measures a unidimensional construct (Paek & Cole, 2019). Several IRT models were used to model the data, such as the Rasch model, 1-parameter logistic model (1PLM), 2-parameter logistic model (2PLM), and 3-parameter logistic model (3PLM) for dichotomous items; Partial credit model (PCM), Generalized partial credit model (GPCM), Graded response model (GRM) for polytomous items. In addition to modelling different types of items separately, nested models were used together to model all the items (such as 2PLM + GPCM). Although these models provided similar results, 2PLM + GPCM was used to analyze all the items together because it provided the best model-data fit and the thresholds estimated by GPCM are easier to interpret (Naumenko, 2014). The R software package was used to analyse the data as it is easy to use and it is free (Hohensinn, 2018). G2 statistics and S-X2 statistics were used to check model-data fit using the *mirt* package in R (Paek & Cole, 2019). Several graphs resulting from the analysis were checked, including the Item Characteristic Curve (ICC) for dichotomous items, Category Characteristic Curves (CCC) for polytomous items, and Test Information Curve (TIC). An ICC or CCC was used to examine the functioning of response options in an item (Paek & Cole, 2019; Bond & Fox, 2015). Test information is the amount of information provided by all the items on the continuum of the measured competence, and well-targeted

persons have more information than do poorly targeted persons (Bond & Fox, 2015). Items with poor characteristics at this stage were deleted or modified.

Despite the similarity between the Rasch model and other IRT models, it has been claimed that Rasch models are “confirmatory and predictive” that require the data to fit the model, while IRT and TST are “exploratory and descriptive” that account for all the data (Bond & Fox, 2015, p. 507). Given the main study aimed to provide an accurate measure for each student and the parsimony of the Rasch model, a Rasch model was used in the main study. Rasch models have been widely used in science education research (Romine et al., 2020; Shi et al., 2021). The Rasch model gives estimates of students’ ability and items’ difficulty on the same scale, and the estimation of one of them is independent of the sample of the other that is used (Wright & Mok, 2000). This therefore allows for exploring the assessed SAC as a learning progression (Osborne et al., 2016). The items in this study were scored either dichotomously or polytomously, so an extension of the Rasch model called the partial credit Rasch model (PCM) was used in this study (Wright & Masters, 1982).

Specifically, the model data fit was examined by:

- 1) testing whether the data follows the assumptions of Rasch model, namely unidimensionality and local independence of the data,
- 2) referring to the fit indices of the mean square residual (MNSQ) to see how much the data aligns with the Rasch model,
- 3) checking the reliability measures. In more detail, the eRm package in the R software was used to model the data (Bond & Fox, 2015).

Rasch factor analysis was conducted by applying PCA (Principal Component Analysis) to the residuals after the primary Rasch dimension has been extracted from the items. The *Pairwise* package in R (Paek & Cole, 2019) was used to perform the PCA analysis. Items that have substantial correlations unexplained by the primary Rasch measure have factor loadings that are greater in magnitude resulting in large eigenvalue of the first principal component. Yen’s (1984)  $Q_3$  statistic was used to check for possible local dependence of items using the *pairwise* package in R.  $Q_3$  is a correlation between item response residuals also accounting for non-linear relationships between the partialled-out primary Rasch dimension and item responses (Yen, 1984). In the Rasch model, reliability is estimated both for persons and for items. The separation reliability is the ratio of the “true” (observed minus error) variance to the obtained

variation with values ranging between 0 and 1 (Duncan et al., 2003; Bond & Fox, 2015).

As in the fourth pilot, several graphs resulting from the PCM analysis were checked including a Wright map. A Wright map displays the location of item parameters and the distribution of person parameters along the latent trait on the same scale, which is used to see how well the item difficulty distribution matches the person ability distribution. It also demonstrates the pattern of item difficulty to reveal the potential structure of the construct that is being assessed (Wilson, 2004). Next item difficulty was examined, which is represented by threshold/item parameter that refers to the point on the underlying trait continuum in which an individual has a probability of 0.50 of selecting a particular response (Bond & Fox, 2015). In terms of the thresholds used for polytomous items, an intersect point (on a logistic scale) represents where there is a 50% probability of being observed in the category below and 50% being observed in the category above the category transition point. The threshold indicators should be ordered from lower values to higher values since higher thresholds require higher ability to get to that score.

Items with poor fit were analysed and decisions were made regarding whether to delete them or modify them (Wilson, 2004). The expected ICC together with the empirical plots (i.e., the actual frequencies of positive responses) were used to diagnose the underfitted dichotomous items. *Empirical* means *observed*, so the empirical function is the actual response function that the item produces, not what the Rasch model assumes. The plot shows whether the Rasch model is accurately predicting how people are responding by exploring whether the two ICCs follow the same trend.

#### **4.5.4 Descriptive statistics**

As mentioned in section 4.4.1.1, a small survey at the end of the SAC test paper was provided to provide information that can be used to evaluate the administration of the test. Additionally, RQ 3 is concerned about the overall picture of the group of Chinese high school students' SAC test performance. Therefore, descriptive statistics was considered helpful to provide frequencies or proportions of each response category in the small survey to help answer RQ 2, and of each performance level in students' test scores to answer part of RQ 3.

#### **4.5.5 Inferential statistics**

The purpose of conducting inferential statistics in this research was to support the interpretation

of the nature of SA as manifested by Chinese high school students in the main study. The set of Rasch scores obtained from the PCM, which is a continuous measure, were used to represent student's performance on the test. RQ 4 is concerned with comparing the SAC performance of students from different subgroups (nominal variables), i.e., area, school, class, gender, with/without scaffold, in order to understand the influence of these variables on SAC performance. Given the nested nature of the data in this study (i.e., students nested within 38 classes), a multilevel modelling approach was considered appropriate because it accounts for the levels of the hierarchy in the population thus enabling researchers to draw more reliable research conclusions (Browne & Rasbash, 2011). Therefore, a two-level students-within-classes variance-components model for SAC Rasch score was firstly conducted by using the MLwiN software to see whether and to what extent the data were clustered within classes (Rasbash et al., 2000). After that, a two-level random-intercept model was employed because there was significant between class variation (see Chapter 7). As only a small number of schools/areas were involved in this study, schools/areas were considered as fixed effects in the model represented by a series of dummy variables to account for the variability caused by school/area difference. Similarly, gender and scaffold variables were added as fixed effects in the model respectively to explore the difference of SAC performance between girls and boys, and between students provided with the scaffold and those that were not.

When exploring the relationships between students' SAC performance, Chinese performance and Physics content knowledge, a correlational analysis was considered appropriate given all these test data were numerical variables. Thus, either Pearson correlation analysis or Spearman correlation analysis was conducted using SPSS based on the normality of the data distribution.

## **4.6 Ethical considerations**

This section will discuss several ethical considerations that were significant to this study and dilemmas faced in practice based on the ethical procedures and regulations set by the School of Education.

### **4.6.1 Researcher access and informed consent**

Participants in this study were between the ages of 16 to 18 and did not include those who are incapable of making their own decision because of immaturity or any psychological impairment (Cohen, Manion, & Morrison, 2002). So, informed consent was gained from involved students and teachers themselves.

For the first three small-scale pilot studies, teachers were contacted personally, and informal consent were obtained from these teachers. For teachers that were contacted to help recruit students, they were explained to about the students' right of participation, the principle of voluntarism of recruitment and their right to withdraw within 10 days after collecting the data. After the formal data collection began, all the participants were provided with the information sheet that explained the research aim and the activities that they were invited to do in PDF format (see Appendix 8 to 10) and the consent form that explained their right of participation (see Appendix 11 and 12) to sign.

However, in the two large scale studies (i.e., fourth pilot and main study), things became complicated since tensions emerged between the gatekeepers (i.e., teachers or principals in schools) and me as a researcher and as an acquaintance/stranger to these gatekeepers. Some of the gatekeepers were my previous colleagues in Master study, while others were known by my colleagues and were invited because of their interest in the project and their generosity to help. They all showed their support to me as a researcher and their recognition to the research project when I contacted them. However, pandemic conditions led to several delays to the start of data collection, which made things became more complicated than expected since we kept changing the date of data collection.

Then, when data collection began, the procedure seemed to be more complicated than expected, as well as the gatekeeper's workload brought by participating in the research. Due to the restriction of COVID-19, I was not allowed to enter the classrooms, so all the materials needed to be distributed to the students by the gatekeepers. Given the large amount of data needed in the fourth pilot and the main study, the number of materials needed to be distributed (information sheet, consent form, test paper), and the fact that there were two types of test papers that needed to be deliberately distributed within classes. The gatekeepers and I both realized the complexity and time that would take. Considering teachers were all quite busy especially high school teachers in China and the delay of data collection, I would not like to add more burden to them. So, I discussed with them about what would be better practices for them and decided to use an alternative way to inform students. Finally, in the fourth pilot, I recorded a video in which I explained the research and the participants' right, the teachers could choose either read the information sheet and consent form to the students or play the video to them. It turned out that not all the teachers prefer to play the video due to the limitation of media player in the classrooms. Similarly, teachers in the main study expressed that it would



be unnecessary and time consuming to provide each student another two documents to look at, so they chose to read it to their students before the assessment and emphasized the volunteer nature of this research by telling the students that they have the right to choose not to engage in the test.

Participants who were willing to participate in the follow up interview signed an abridged version of the consent form on the end of the test papers (see Appendix 19 and 21). I emphasized the voluntary nature of the research to them again after I contacted them. The follow up interview showed that students knew that the test was voluntary and there were as a result some students who chose not to take the test (or not to submit it). But given teacher's authority in school, it is still possible that some students may have taken the test due to perceived pressure from their teacher.

#### **4.6.2 Anonymity**

Considering the research design of this study, students were invited to fill in their name on the test paper, but they also gave the right to be anonymous. The reason why it was not anonymous is that in RQ 4, students' school test scores were compared to the SAC test scores, so there was a need to match each student's scores on the two tests. However, after matching students' school test and SAC test, their names were removed from the data prior to analysis. In the following coding and analysis process, all the data were treated confidentially and anonymously. Only researchers and supervisors have access to the data.

#### **4.6.3 Participants' right and experience**

Students and teachers in high schools in China were under great pressure of the Gaokao at the time of this study and they had very limited time to do other unrelated things in general. So, participants were emphasized of their right to withdraw from taking the test or the interview. Additionally, considering the instrument was a test, students might feel bad if they found it difficult. Teachers were asked to emphasize to students that it was not a high-stakes test, and the assessment results would not be taken as an indicator of their academic level but used for research purposes. Furthermore, to reduce the possibility of students not taking the test seriously, teachers were asked to tell students the importance of their presentation of real ability, and the possible contribution this study would bring to educational practice. However, this turned out to bring another potential limitation of the study, that is teachers in different classes/schools treated it differently with some teachers conveying a more serious signal to

students and presented in the classroom to supervise the test, while others did not (this is known by asking the students during interview). Additionally, the voluntary nature made it the fact that not all students in a class consented to take the test, which may have influenced the inferential statistics analysis that will be discussed in Chapter 7.

Students in the interview were fully respected and I tried to make them feel comfortable and that they were being helpful by participating in the research. For instance, the interview time and the way of interview (i.e., voice call or video call) were all based on their time schedule. During the interview, I tried to build a rapport with the participants by asking questions like ‘How was your day?’ and ‘Are you busy today?’ etc. In addition to asking what I was interested, I also paid close attention to what they wanted to share without interrupting them from sharing. I answered each of their question (many of them are irrelevant to this research, such as personal issues and their concerns about their future etc.) sincerely and with patience. For instance, some students’ interview may last for 2 hours since they were talking with me about other issues they encounter, but these contents were not recorded. For students who showed concern about learning English or other subjects, I shared some learning websites or materials with them after the interview. They were told that they would be welcome anytime if they want to talk to me. It turned out that several students contacted me several times even after nearly two years of collecting data. Additionally, I responded to my participants in an encouraging way even if they provided incorrect responses.

Some participants mentioned that they felt nervous before the interview began since they had never participated such interview before, but they felt the experience relax and happy after the interview began. Some students also mentioned that they might feel more nervous if it was face-to-face interview. Although voice call made interviews convenient for the participants, teachers, and me, some students could not talk due to some reasons and I could hear some obvious background sound (i.e., people talking) when I had interview with few participants. I could sense that there were some external factors that influenced their engagement and engagement (i.e., a tendency to remain silent during interviews). So, instead of forcing them to speak, I conducted brief interviews with these students.

All the participants were provided with reward personally (for participating in interview) or for the whole class (for taking test). For the participants in pilot I, I bought some chocolate from the UK and brought to them when I went back to China in 2020. For the participants in pilot III, I bought USB flash disks for them. For participants in pilot II and pilot IV and the main

study, each class were given 50 RMB (about 7 GBP) or 100 RMB (about 13 GBP) as their class bursary, and students who participated in the interview were given 20 RMB (about 2.3 GBP) each.

### **Chapter summary**

This chapter justified why pragmatism highly fits into this study and makes it appropriate for this study to employ a mixed methods methodology and an iterative research design. Specifically, four pilot studies were conducted with the aim of developing and modifying the SAC assessment, with three pilots focusing on think aloud interviews and follow-up interviews and the fourth pilot involving large-scale test and follow-up interviews. The data in each pilot were analyzed to inform the SAC assessment design used in the next pilot and then in a main study. The data analysis in the pilot studies together informed the validation of the SAC assessment and the design of SA assessments.

A main study was conducted to obtain large-scale information of the final version of the SAC assessment and to explore the students' experience related to SA by administering large-scale test and follow-up interviews. Background information were obtained in the main study for statistical analysis to capture an overview of how class, school, gender, assessment scaffold, Physics and Chinese achievement relate to SAC performance. Ethical dilemmas encountered in the research especially informed consent were reflected upon to inform the limitation of the research.

Overall, this chapter provided an overview of how the research was designed in relation to addressing the research questions. The next chapter will discuss what the design and modification of the SAC assessment in the iterative research design looks like and how results obtained from the iterative process could inform future SA assessments.

## **Chapter 5. Developing a SAC Assessment**

### **Introduction**

This chapter aims to answer Research Question 1 concerning the assessment design. The framework used to guide the assessment development will be introduced in section 5.1. Four versions of the assessment instrument will be explicated in subsequent sections, highlighting how each design was modified based on empirical findings from pilot studies.

Overall, this chapter will make open and transparent the development procedure of the SAC assessment, the design process and its products, and the strategies employed during this process to improve the assessment design, which may be worth considering when designing other SA assessments. Therefore, this chapter lays the foundation for achieving the research aim of assessing and understanding Chinese high school students' SAC.

### **5.1 The four building blocks for constructing measurement**

As argued in section 3.5, it is as important to justify the process of developing an assessment as to scrutinize the product. Thus, it is sound and resource-saving to guide the process with existing frameworks for developing assessments given the principles and practices of educational measurement have been well-established. This study mainly draws upon Wilson's (2004) approach of constructing an assessment because it focuses on the development of a construct rather than only of content knowledge, which fits the intention of exploring SA as a learning progression. Additionally, Wilson's approach includes four building blocks that are in accordance with the development stages, thus, using his framework to guide the assessment development helps generate information in terms of the process of developing the assessment. Certainly however, some ideas from other resources were also applied in this study. For example, texts like "Standards for Educational and Psychological Testing" (AERA, 2018) and "Handbook of Test Development" (Lane et al., 2015) provide detailed insights into the development, delivery, and analysis of a variety of assessment formats. However, the discussion in these texts is mainly dominated by large-scale multiple-choice item tests that are usually used in school-achievement tests. Each building block of Wilson's approach is introduced below.

### 5.1.1 Construct map

It is the first step in assessment to articulate what is to be assessed and how it develops, which serves as a guide to item writing. The concept *construct* here refers to the object to be measured by the instrument, such as ability, understanding or attitude. *Construct map* is a graphical representation describing how the *construct* develops from one extreme to another, such as from low to high, from simple to complex, and from weak to strong (Wilson, 2013). The idea of a *Construct map* but under different names has been widely applied by other researchers (Wyse, 2013). Construct maps are derived in part from research and professional judgments about what constitutes higher and lower levels of competence and might be modified based on empirical evidence of students' performance in practice (Wilson, 2009; Wyse, 2013). For the aim of measurement, what matters is not the complex structure between the two extremes, but the location where a respondent stands on this continuum. Although SAC cannot be observed directly, the underlying continuum can be manifested by the ordered levels of respondents or their responses.

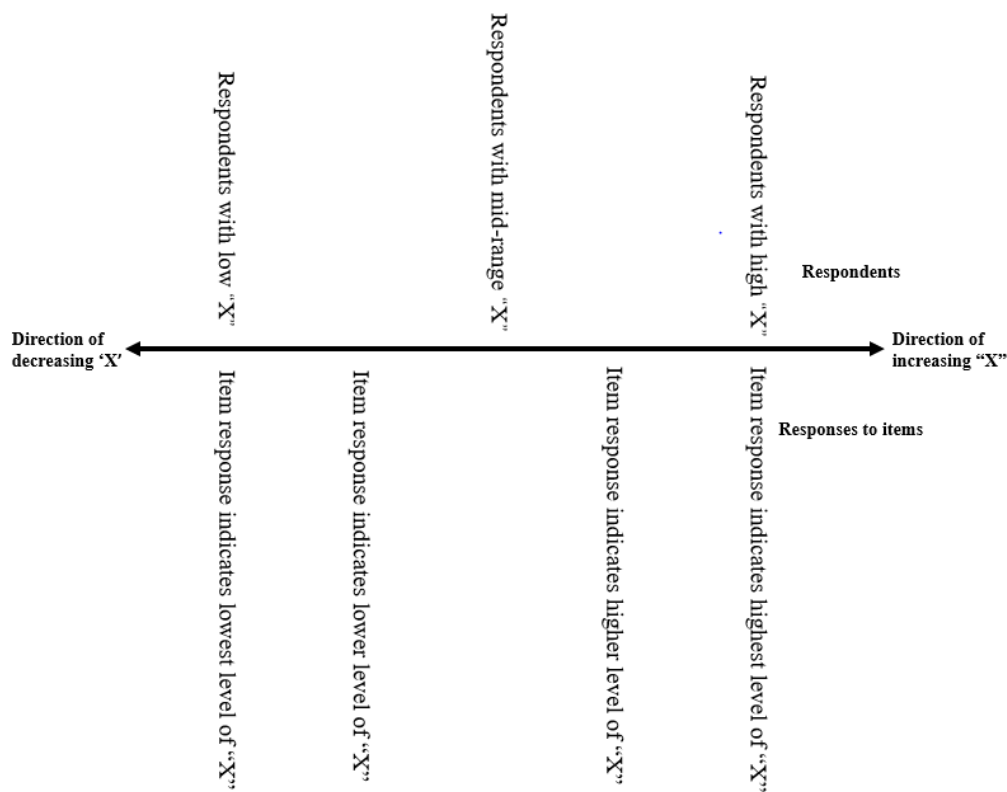


Figure 5.1 A generic construct map in construct "X" (Wilson, 2004)

### 5.1.2 Items design

*Items* can be taken as a medium through which a theoretical construct can be manifested empirically, and *items design* refers to the types and nature of items to be used in the measurement (Wilson, 2013). Items should be developed to target the skills and competences articulated in the *construct map* and can be represented in different formats, such as multiple-choice questions and short answers. Item design reflects a series of decisions made by the item designer, such as whether it is a self-report item or performance item, their relationship with the curriculum, actions needed for respondents to complete the item, etc. These decisions might be influenced by the nature of the construct, practical considerations, and even arbitrary choices of the designer. The characteristics of a set of items should be explicitly described in the early stage of instrument development even though they may be modified during the whole development process. Otherwise, if the items are designed tentatively before the item designer has any specification of the whole item design system, then the initial items can be seen as part of the *item design* development process.

Interviews, observations, literature reviews and initial drafts of items could all contribute information about what to focus on and how to express the questions appropriately when developing an instrument. Obtaining information from respondents is a significant step in designing and improving an instrument. There are mainly two ways to investigate information from respondents: think aloud processes and the exit interview. Think aloud processes aim to obtain information about students' response processes (i.e., the process of figuring out test items) while they attempt the items, while exit interviews aim to investigate student's response processes and reflections after they have made responses. Except for information from respondents, feedback can also be obtained from teachers or professionals in the content area. So, item paneling composed of several people who are professionals or knowledgeable and reflective about the area of interest can be conducted during the item development and revision processes. The above idea of designing items has informed the research design and data collection methods of this study, as presented in section 4.2 and section 4.4.

### 5.1.3 Outcome space

*Outcome space* is a set of qualitatively described categories of responses to a task, which relies on qualitative understanding of what constitutes different levels of response (Wilson, 2013). The outcome space for an item is used to categorize results, such as in multiple-choice items,

1 refers to correct and 0 refers to wrong. The characteristics of the outcome space are well defined, finite, and exhaustive, ordered, context specific, and research based (Wilson, 2004, p. 62). Like the item design process discussed in the previous section, the construction of the outcome space is also an iterative process that needs to be revised across the instrument development period.

#### **5.1.4 Measurement model**

The *Measurement model* refers to the way of relating the scored outcomes and the outcome space back to the construct that it was intended to assess (Wilson, 2013). Examples include the *Rasch model* based on Rasch's (1960) work and similar research termed *Item response theory*. It is important to relate the scored data back to the *construct map* to understand both the construct and the instrument using the measurement model when developing the instrument. A measurement model is used after an assessment has been designed and administered. The idea of using IRT models to analyze the test data has been talked about in section 4.5.3.2.

The above four blocks constitute the whole process of developing an assessment, each block can be modified until an acceptable assessment is generated. The subsequent sections will introduce how the blocks of the SAC assessment were designed and modified.

#### **5.2 Assessment design I- An initial attempt**

The procedure of the iterative instrument development process is shown in Figure 5.2. This figure is similar with the one in section 4.2 but with a focus on the modification of the instrument. Before collecting data, the initial assessment instrument that included an initial construct map (i.e., Construct map I), several initial items (i.e., Test version I) and corresponding initial outcome space (i.e., Scoring rubrics I) was designed based on the extant literature and the aim of this research. As an exploration of achieving the aim of assessing SAC from three components (i.e., I-SA, E-SA, and P-SA), the initial items were seen as an entry point to the *Items design* process as pointed out by Wilson's (2004) approach.

By conducting pilot I, the initial assessment was modified to generate an updated version of assessment (i.e., including Construct map II, Test version II, and Scoring rubrics II) that was used in pilot II and the subsequent iterations were implemented similarly. Each version of the instrument will be explicated following the 'four building blocks' framework subsequently. All versions of the test are translated into English here simply for the convenience of readers.

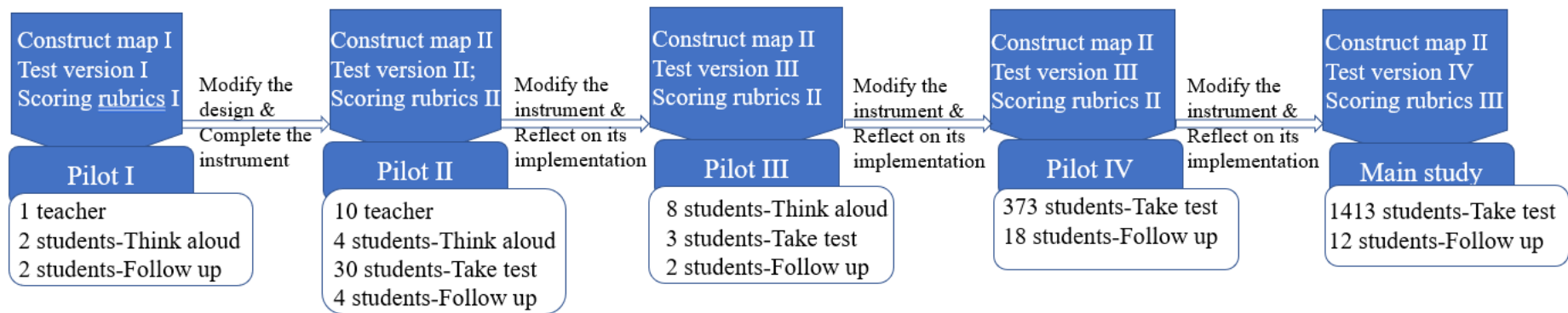


Figure 5.2 Instrument development procedure



### 5.2.1 Construct map I

As has been identified in section 3.2, this study understood the construct of SAC as containing three components: Identifying a scientific argument (**I-SA**), Evaluating a scientific argument (**E-SA**), and Producing a scientific argument (**P-SA**). The structure of SA was further deconstructed based on previous studies.

Chapter 3 has discussed the identified difficulty of differentiating between the elements such as ‘warrant’ and ‘backing’ (Duschl, 2008; Erduran et al., 2004; Jiménez-Aleixandre et al., 2000), this study therefore modified Toulmin’s argument pattern (TAP) to structure *the argument product* generated during the process of SA as containing *claim, evidence, reason, and rebuttal*. A *claim* is an assertion or conclusion people believe in; *evidence* is the data that can be used to support the claim; *reason*, although represented by different terms across the literature (Osborne et al., 2016; Erduran et al., 2004; Sengul et al., 2020), is the connection between claim and evidence explaining why the evidence supports the claim; *rebuttal* is a statement put forward to undermine an argument. In contrast, drawing upon Sandoval’s (2003, 2005) study on scientific explanation (see section 3.3.3), this study perceived the *operation of constructing an argument* as including *use of evidence, explanation, and rebuttal*, that is a person should be able to use evidence, explain and generate rebuttal in order to be engaged in SA.

Therefore, I-SA referred to student’s ability to identify the function each statement plays in a piece of argument, focusing on the structure of SA. Specifically, students should be able to *identify* the different elements (claim-Ic, evidence-Ie, reason-Ir, and rebuttal-Irb) in an argument. In a similar way to Osborne et al.’s (2016) study, which takes student’s ability to identify argument elements as the lower levels of critique of argumentation, this study perceived it as a component of the SAC that reflects students’ understanding of SA and the very first step of engaging in SA explicitly.

E-SA represents student’s ability to evaluate the quality of SA and P-SA tests whether students could generate the elements in an argument. However, the initial focus of creating an SAC construct map was on deconstructing SAC into sub-competences and expose each of them explicitly, so E-SA and P-SA aimed to consider the *operations of constructing an argument*, namely *use of evidence, explanation, and rebuttal*. As will be shown in section 5.2.2, the *Evaluation* and *Production* of these operations were assessed separately as sub-competences

without being put in a context of SA, and it was assumed that if a student possesses these sub-competences, he/she would be able to engage in SA. To put it another way, it was assumed that if a person can operate changing gears, hitting the brake, stepping on the gas, etc. he/she would be able to drive a car.

In fact, the evaluation of SA is complex, not only for various aspects of criteria but for the degree of what is defined as ‘good’. This study, as an exploratory study on this aspect, focused on several main points for each element of SA rather than extending it in a complex way. Putting it simply, this study listed a few criteria for each item to obtain a general picture of whether students can match the criteria with a given piece of argumentation. The criteria of E-SA items drew on previous theories and studies on SA as reviewed in Chapter 3. The criteria for *Evaluating use-of-evidence* were **relevance** and **sufficiency**, which cares about

- 1) whether the evidence is relevant to the discussed issue, and
- 2) whether the evidence is sufficient at supporting the claim (Rapanta et al., 2013).

The criteria for *Evaluating explanation* were **causal relationship** and **coherence**, which emphasizes

- 1) whether the reason provides the causal relationship between evidence and claim, and
- 2) whether the explanation is coherent (Sandoval, 2003, 2005).

The criteria for *Evaluating rebuttal* were **accuracy** and **reasonability**, which focuses on

- 1) whether the rebuttal is aimed at the problematic point of the other person’s argument accurately, and
- 2) whether the provided rebuttal is reasonable.

As discussed in section 5.1, there is always an underlying continuum to manifest the construct. Previous studies asserted that critique is more difficult for students than constructing an argument (Osborne et al., 2016; Erduran et al., 2004). Evaluation can sometimes overlap with critique, as implied in Osborne et al.’s (2016) study. However, the criteria have been given to the students and the students in this study do not need to explicate their evaluation or to compare different arguments to decide which is better for addressing an issue / answer a question. And as introduced in section 3.4, students were found to be able to generate argument from an early age, thus generating simple argument should be easy for high school students.

Therefore, it was hypothesized that I-SA should be the easiest, followed by E-SA and P-SA in general, but their detailed order of difficulty needs empirical investigation. As for SA elements, based on previous studies as shown in sections 3.3 and 3.4, it was hypothesized that rebuttal is the most complex, followed by reason and evidence. A graphical representation is shown in Table 5.1.

*Table 5.1 Construct map I*

SAC component	SA element	Direction of complexity (from low to high)	
Identification of SA	Claim Evidence Reason Rebuttal	↓	↓
Evaluation of SA	Use of evidence Explanation Rebuttal	↓	
Production of SA	Use of evidence Explanation Rebuttal	↓	

### 5.2.2 Test version I

The items design at this stage was not done in a systematic and coherent way but was used to explore potential questions to which students can apply the sub-skills of SAC. As an exploration of designing SAC items, items covered two content topics in high school Physics: Motion and Force, and Electricity. The initial principle for this step was to choose topics in the curriculum that grade 11 students had learned. The remaining decision about content topics depended on how easy and appropriate it was to design a SAC item. As for the level of difficulty, the Physics curriculum was used as a reference to decide which aspect of knowledge might be suitable for the aim of this assessment. The overall consideration was that content knowledge should be at a fundamental level in the curriculum. Besides, not all the SAC elements had corresponding items being designed at this stage. There were 6 tasks assessing respectively: I-SA, E-SA-use of evidence, E-SA-explanation, P-SA-explanation, P-SA-rebuttal, P-SA (see Appendix 14).

The feature of these items was that each item used a single scenario to assess a certain SAC element, so that

- 1) it is easier to match a scenario with certain sub-competence and
- 2) items assessing different sub-competence would be independent (a requirement of IRT)

model) and the items that representing different SAC elements could be identified as a progression.

I-SA elements were assessed using the same scenario to save resources given it was assumed to be the easiest SAC component. Whilst a sixth task contained three items that assessed the three elements of P-SA in turn to see how these two task forms might work differently. Overall, this version of items paid more attention to each SAC element as a separate task. Such as the P-SA-explanation task shown in Figure 5.3, which assessed student's ability of explaining.

**Task 3.**

Table 1 records the weight of a student (weighing 55kg on the ground) at four moments when taking an elevator.

*Table 1*

T 1	T 2	T 3	T 4
55kg	50kg	55kg	60kg

What kind of movement might the elevator undergo? Why?

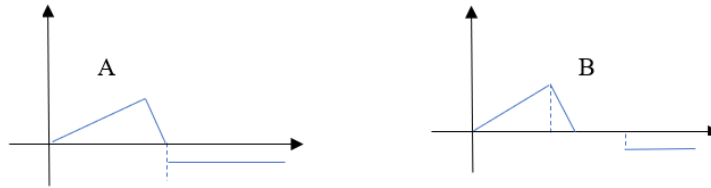
*Figure 5.3 P-SA-Explanation task example (Test version I)*

As for item format, all the I-SA and E-SA items were multiple-choice or match questions, and the P-SA items were open-ended questions. This decision was a result of a balance between practical consideration and the predefined theoretical nature of SAC components. It should be easier to find out students' thinking processes if asking the students to *Evaluate* argumentation using open-ended questions, but the time for taking the test and scoring would increase as well. The second reason of providing the criteria for students to choose rather than asking them to evaluate SA directly using their own criteria was to approach students' competence gradually and pedagogically especially considering Chinese students' unfamiliarity with SA. The last reason was that this study intended to explore the extent to which Chinese high school students can evaluate SA based on the norms of argumentation.

An I-SA task is shown in Figure 5.4, in which Xiao Li's argument about a question is provided and students need to identify the four SA elements in his argument. The content knowledge is using a graph to solve motion problems, which is fundamental in high school Physics.

**Task 1.**

After receiving a delivery phone call, Xiao Li rushed to the pickup point and then walked home after picking up a package. There is a straight road between the delivery point and home. A and B are possible  $v-t$  images of Xiao Li's movement from going to pick up the package to returning home.



Xiao Li agrees with the process drawn by Figure B:

- (1) The first picture does not indicate the movement status of waiting for pickup.
- (2) The beginning of the second graph is a straight line with a positive slope, and then a straight line with a negative slope. Both are on the positive semi-axis of the y-axis. The next section coincides with the horizontal axis, and the last horizontal straight line is on the negative semi-axis of the y-axis.
- (3) Xiao Li runs for the delivery, so speed up the movement, and decelerate before reaching the pickup point. Walk home, so the speed will be smaller, and the direction of movement is the opposite. In the  $v-t$  image, a straight line with a positive slope represents uniform acceleration, a negative slope means uniform deceleration, a horizontal straight line means uniform speed, and coincidence with the X axis means that the speed is 0.
- (4) The second picture is the most reasonable.

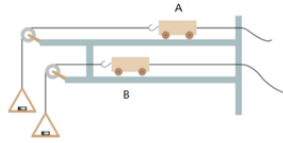
In the above description, Xiao Li's claim is \_\_\_\_\_. The evidence he used to support his claim is \_\_\_\_\_. His reason for using the evidence is \_\_\_\_\_. His rebuttal towards the first picture is \_\_\_\_\_.

*Figure 5.4 I-SA task example (Test version I)*

Figure 5.5 below is an E-SA task that includes 3 items assessing student's understanding toward *Use of evidence* with each item focusing on one criterion. In this item, evidence was not designed based on facts, but on data from students' experimental results. Thus, evidence b and c are contradictory in nature, and that's why one more criterion was added indicating content knowledge (although both items of evidence are relevant, only one can support the claim after analyzing the question using relevant content knowledge). Students needed to match the given statement of using evidence with the three criteria provided in the task.

**Task 2.**

As shown in the figure, cart A and B start to move from standstill under the pulling force of the groove plate. They move for the same time (smooth track, smooth pulley). Student a, b and c each gives evidences to support the conclusion that “the mass of the slot code in slot B is greater than that in slot A”.



a: Cart A uses more environmentally friendly materials than Cart B, and Cart B uses thicker ropes than Cart A.

b: Both carts have the same mass, and the displacement of A is greater than B.

c: Both carts passed the same displacement.

Student d thinks that the evidence given by other students cannot support the conclusion, please choose the reason in ①②③ to fill in the blanket.

a ----- reason: (    ); b ----- reason: (    ); c ----- reason: (    )

- ① Evidence contradicts the conclusions.
- ② Evidence is not relevant to the conclusions.
- ③ Evidence is insufficient.

*Figure 5.5 E-SA-use of evidence task example (Test version I)*

**5.2.3 Scoring rubrics I**

Different levels of scores correspond to the various possible answers from students and aim to differentiate students’ performance on the task. The rubrics shown below are for Task 3 (see Figure 5.3). The grain size (i.e., the level of detail or the level of difference between adjacent responses) of this initial rubric was quite small since many possible answers were considered in order to cover all the potential answers from the participants (see Appendix 15).

*Table 5.2 Scoring rubric example (Test version I- task 3)*

Task 3: Production Item 3-P-EX: Generating explanation	
Score	Description
5	Student uses coherent articulation to explain the process with correct causal relationship.
4	Student provides correct causal relationships and understand why this phenomenon happens without explaining the process or explains it not fully correct.
3	Student tries to provide explanation, but conclusions or causal relationships are incorrect.
2	Student does not give explanation but provides correct conclusion to the process of movement.
1	Student does not give explanation but provides partly correct conclusion to the process of movement.
0	Student does not give explanation and does not provide correct conclusion to any points or stage of movement.

### 5.3 Assessment design II- A rich exploration

The main finding of pilot I was that students did not seem to engage in an argumentation. So, the design of Test version II aimed to create an atmosphere of argumentation to better elicit students' SAC. In general, Assessment design II, was designed in a more systematic way, and was a complete version that included all the required items. Assessment design II will be elaborated upon by illustrating how the findings from pilot I contributed to it.

#### 5.3.1 Construct map II

Students in pilot I tended to focus on using formulas to make sense of the provided phenomenon without being engaged in SA (i.e., thinking about evidence and the connection between evidence and claim) when dealing with the P-SA-Explanation and E-SA-Explanation items. This raised reflection in this study about the long-lasting debate between the relationship of scientific argumentation and scientific explanation as to whether they are closely intertwined or should be treated separately (Osborne & Patterson, 2011, 2012; Berland & McNeill, 2012; Berland & Reiser, 2009; Brigandt, 2016). So, a closer look at the construct map was needed.

The process of framing the assessed competence and designing items that test it are always reciprocal. As described in section 5.2.1, each scenario was used to assess only one E-SA or P-SA element in the initial items design. For example, instead of integrating *Explanation* into a context of argumentation, P-SA-Explanation was assessed using the lift scenario asking students to explain why the number on a weighing scale changes as the lift moves (see Figure 5.3). There should be no problem of including *Explanation* into the construct given the literature review of SA and SA research, but the exact meaning of *Explanation* then needs to be clarified. To put it another way, the thinking process or the activity of *Explanation* that happens under different contexts has a different nature, although its central meaning is *to explain*.

So, when defining or assessing a construct, the context under which it happens should also be made clear. In this case, *Explanation* should happen under the context of argumentation rather than on its own. Thus, the design of assessing *Explanation* separately from the activity of *Argumentation* brought the risk of assessing another construct although their meaning seems the same. Therefore, this study decided to replace *Explanation* with *Reason*, highlighting that what matters in the assessment was not its general meaning of *answering the question 'Why?'* (Osborne & Patterson, 2011), but the meaning of *explaining why certain evidence can support*

*certain claims in an environment of argumentation.*

The construct map of SAC therefore was modified in a more specified and clearer way (see Table 5.3). This study finally perceived SA (as a product in a specific context) as including *claim, evidence, reason, and rebuttal toward a specific problem*. Whereas perceiving the operations of constructing SA as including **providing evidence and reason toward the proposed claim of the specific problem** and **providing rebuttal towards the statements related to the specific problem**. In particular, *reason* represents **the explanation of the connection between the evidence and the claim of a specific problem**. Therefore, different from in Construct map I, it was assumed that if a student is able to engage in SA, he/she should be able to show these sub-competences **in the context of SA**. Similarly, if a person can drive a car, he/she should be able to operate changing gears, hitting the brake, and stepping on the gas etc. **in the context of driving**.

These SAC components and elements together depict the competences students needed to engage in scientific argumentation that encompasses both “structural and dialogic” focus (González-Howard & McNeill, 2020, p. 957).

*Table 5.3 Construct map II*

SAC component	SA element	Direction of complexity (from low to high)	
I-SA	Claim Evidence Reason Rebuttal	↓	↓
E-SA	Evidence Reason Rebuttal		↓
P-SA	Evidence Reason Rebuttal		↓

### 5.3.2 Test version II

Test version II was created by modifying Test version I based on the findings from conducting pilot I. So, the main findings from pilot I that guided the creation of Test version II will be introduced first as below.



### *5.3.2.1 Main findings from pilot I*

Four themes were constructed by analysing interview data in pilot I, namely **‘Limited engagement in SA’**, **‘Language’**, **‘Scenario design’**, and **‘Familiarity with SA’**. The teacher (F1) (this is an identifier for the participant) expressed her feeling that the test was more about content knowledge rather than SA, and she pointed out that Task 4 (using a shared scenario, see Appendix 14) would be too difficult for students because students have never met the scenario in school learning and the information provided in the task was too much. She also pointed out that the involved content knowledge was difficult for most students: “even these content knowledges are difficult for most students, not to mention the high order thinking skills”.

Additionally, students’ interpretation of language seemed to be influenced by their previous experience of learning in the school context. For instance, when they were asked “who do you agree with?”, they tended to choose one side to agree with even when they agreed with neither of the sides. Three out of the 4 participants misunderstood the question and interpreted it as: “Please choose one to agree with” (FF3). Similarly, student FF1 said that: “I was thinking that neither of them is right, but the item question asks me to select one to agree with, so I selected one randomly.”

In addition, in order to set the scenario close to real life, more words were presented in item stems to make it understandable. But it did not work out as expected. Students often found it even harder to understand or tended to miss some information since they lost patience when reading the statement word by word. For example, students who participated in the think aloud interview all got confused about Task 1 (see Appendix 16) and spent a long time to understand the item stem. Student FT4 said, “I really get confused, it is such a long story to me” and FT2 said, “the expression here is really confusing”.

As introduced in section 5.2.2, most of the items were designed so that each scenario assesses one element of SAC in the initial item design. However, this design created more workload for students since they needed to deal with too many different problems, and students spent lots of time even on just a few tasks that did not cover all the elements of the construct.

Students mentioned that they had never before been engaged in SA and knew little about it, and they were not sure about what the test was testing. Moreover, it appeared that students were thinking more about content knowledge without engaging in the process of argumentation

although they felt the items “need more thinking and explanation” (FF1). For instance, FT4 mentioned that “I forgot about the conclusion, my teacher has told us about it” and FF1 said “I know the results but forgot about why, I didn’t know how to explain it” when responding to the item in Figure 5.3. This not only shows that the item did not highlight the feature of SA, but also indicates that it provided more possibilities for students to recall the conclusion (the item was adapted from the textbook, and the students were familiar with it). Nevertheless, in contrast from the teacher’s viewpoints, generally, students found Task 4 was more engaging because it is “close to real life” (FT2) although they found it difficult because “most of us found topics related to electricity are often more difficult” (FF3) and “need to deal with too much information” (FT4).

Overall, the reason for students’ not engaging in SA was summarized as

- 1) the wordy language or the imprecise expression confused students;
- 2) the scenario that often appear in their examination papers made them recall conclusion;
- 3) the single scenario design separated SA elements therefore segmented students’ SA thinking;
- 4) unfamiliarity with SA and the requirement on content knowledge impeded students’ engagement.

### **5.3.2.2 Test modification**

Based on the above findings, three strategies were used to improve the test, namely **editing language, changing scenario arrangement, and providing basic information about SA**. Combining these with reflections on the test design, Test version II was updated mainly in the following aspects (see for example Figure 5.6).

The language used in the test was expressed deliberately in a succinct and direct way and was revised across each pilot studies. Also, more scenarios that are close to life and have the potential to arouse argumentation were added (rather than recalling knowledge when facing scenarios that students usually meet in their normal school test) to the test. Each scenario focused on one topic and was linked together using a story line describing a series of scenes two students experience while travelling. Argumentation happens because of the intrinsic uncertainty that exists in the interaction where people hold different ideas (Chen & Qiao, 2020), so dialogues that discuss the targeted topic were provided at the beginning of each scenario to create an environment of argumentation. Items assessing different SAC elements were included

in the same scenario so that these correlated sub-skills won't be segmented by different scenarios. Items under each scenario were ordered progressively (i.e., I-SA items followed by E-SA items and P-SA items), allowing students to be more engaged in the procedure of argumentation.

The content knowledge involved in this test version were changed to be all about Motion and Force to exclude the influence of content topic and as it was a more fundamental topic in High school Physics compared to Electricity. Each SAC elements had at least three corresponding items for future item selection and modification. In total, there were 7 tasks (i.e., scenarios) and 61 items in the test. Except for the supports manifested in the scenario, an argument example and explanation of the SA elements entailed in it were presented at the start of the test and easier content knowledge was involved in the items. Combining with the discussions in section 3.4.2, instructions were found helpful in improving students' SA engagement, thus the argument example and its explanation was taken as an assessment scaffold.

In terms of the options for E-SA items, the item option for *Evaluation of evidence* (Ee) item was

- 1) whether the evidence is relevant and
- 2) whether the evidence is sufficient.

For the *Evaluation of reason* (Er) items, the item options were specified as

- 1) relevance,
- 2) reasonable and
- 3) comprehensive.

The nature of the criteria did not change but were presented in a different way to make it more understandable for students and aligned with the scenario. *Relevance* is whether the reason provided is correlated with the evidence and claim. *Reasonable* is whether the connection between evidence and claim is reasonable, which is an alternative expression of the *causal relationship*. *Comprehensive* is whether the provided reason is coherent enough to fully connect between evidence and claim, which is a replacement of the *coherence* criterion. The expression of the *Evaluation of Rebuttal* (Erb) criteria was more specific as

- 1) whether the rebuttal points out other's weakness,

- 2) whether the rebuttal is based on evidence and
- 3) whether the rebuttal is reasonable.

Each of the options represent one criterion of the SA element, and they were not mutually exclusive to each other for an SA element may meet one or more criteria. So, one or more options could be selected in each item. An example of the test task is shown in Figure 5.6.

**Scene 2: They finally catch the train before it leaves. There were several children in the carriage, and one of them flew his toy helicopter.**



*Bob: It's dangerous to play in the carriage.*

*Jane: Why?*

*Bob: The helicopter will not continue to hover steadily after driving. It will bump into people or the door of the carriage.*

*Jane: No, neither we nor our bags slid back after driving. The movement of stuff in the carriage and the carriage were the same.*

1. What is Jane's claim? It is not dangerous to play helicopter on the train
  - A. There is no claim. B Her claim is \_\_\_\_\_
2. What is Jane's evidence? Neither we nor our bags slid back after driving
  - A. There is no evidence. B. Her evidence is (mark off use "\_\_\_\_\_")
3. If so, which of the following do you think is true of her evidence? (can choose more than one) A
  - A. The evidence and claim are relevant.
  - B. The evidence is sufficient.
  - C. None of above
4. What is Jane's reason? The movement of stuff in the carriage and the carriage were the same
  - A. There is no reason. B. Her reason is (mark off use "-----")
5. If so, which of the following do you think is true of Jane's reason? (can choose more than one) A
  - A. Reasons are related to evidence and claim
  - B. Reasons are reasonable
  - C. Reasons are comprehensive
  - D. None of above
6. Which sentence is rebuttal? No... the same
  - A. There is no rebuttal. B. The rebuttal is (mark off use "~~~~~")
7. If so, which of the following do you think is true of the rebuttal? (can choose more than one) D
  - A. Accurately pointed out the other party's mistake
  - B. Rebuttal based on appropriate evidence
  - C. Rebuttal is reasonable
  - D. None of above
8. Who do you agree with more?
  - A. Bob B. Jane C. My own opinion \_\_\_\_\_
9. Which of the following facts/theories support Bob/Jane/yourself? A, C
  - A. Bodies that are not subjected to external forces have a tendency to remain in motion
  - B. When a train is traveling in a straight line at a constant speed, the object in the carriage is traveling at the same speed as the train
  - C. The train accelerates when it starts
  - D. People lean back when the train is speeding up
  - E. when the train accelerates, the front carriage drives the rear carriage
10. How does the evidence support Bob/Jane/yourself?
11. Why do you think Jane/Bob/both is wrong?

Figure 5.6 Task example (Test version II)

### 5.3.3 Scoring rubrics II

Findings from pilot I showed that some of the responses in scoring rubrics I didn't appear in students' responses. At the same time, it was realized that the very detailed while unsystematic design of the rubric would increase the workload for the rater, which could lead to inconsistent ratings. The updated scoring rubrics for Test version II drew upon Biggs (1982)'s Structure of the Observed Learning Outcome (SOLO) that categorizes students' understanding from surface to deeper thereby their responses varied from irrelevant to relevant but also from simple to comprehensive and coherent (Bowen, 2017). Thus, scoring rubrics was organized in a progressive way with levels indicating more complex responses, that is, the improvement of the skill. The rubric was then almost the same across items that assessing the same SA element (see Table 5.4 and Table 5.5).

*Table 5.4 Scoring rubric example for Pr items*

<b>Score</b>	<b>Description</b>
3	Provide coherent and reasonable reason to connect claim and evidence sufficiently
2	Provide reasonable reason while insufficient; or coherent reason while with flawed content knowledge
1	Provide nebulous reason, with the feature of being illogical, incorrect and without connecting between claim and evidence
0	Does not provide reason, provide irrelevant information

*Table 5.5 Scoring rubric example for Prb items*

<b>Score</b>	<b>Description</b>
3	Provide coherent and reasonable rebuttal to weaken others
2	Provide information that can weaken other's argument, but insufficient or not coherent
1	Does not pay attention to other's argument; or does not weaken other's argument
0	Does not provide rebuttal, provide irrelevant information

### 5.4 Test version III- Abridged and focused

Findings from pilot II did not reveal further problems on the Construct map and Scoring rubrics but indicated that students still lacked awareness and focus on SA engagement. Thus, the aim of Test version III was to focus their attention on engaging in SA.

#### 5.4.1 Main findings from pilot II

Eight themes were constructed by analyzing interviews in pilot II, namely ‘**Better engagement in SA**’, ‘**Test length**’, ‘**Understanding about the problem being argued**’, ‘**Scenario design**’, ‘**Item dependence**’, ‘**SA-related terms**’, ‘**Language**’, and ‘**Content knowledge**’. After modifying the test, students in pilot II showed more engagement in SA. They started to feel that “my mind was focused on thinking about the problem rather than just answering the questions” (ST2) and they were analyzing other’s argument in the dialogue as well, like SF1 said, “I was analyzing what Jane said and I found it is quite reasonable, and I started to reflect on my own claim”. The scoring became easier, and students’ responses were well covered by the different levels of responses in the rubrics. However, more factors that influenced the test were uncovered, which were mainly about the students’ focus on SA engagement.

Firstly, several participants expressed that they got bored when doing the test since there were too many items and the same kind of items repeated in several scenarios, such as the I-SA items. For example, S1 (teacher) said that “I found all the items are almost the same in different tasks...I felt tired reading them again and again.” and S2 (teacher) said “it is too long, my suggestion is to cut off some repeated items”.

Secondly, it was quite common in Test version II that students were first asked “What is Bob’s reason?”, and then “Which of the following is true of his reason?”. Teacher participants pointed out that this design could reduce test validity because if a student answered the first question incorrectly, the second question made no sense. Apart from this, students’ interviews revealed some latent problems on item dependence. Too many I-SA items under one task led students to guess item answers because their misidentification of one element affected their identification of other elements. For instance, SF4 said, “I selected this sentence as reason since there is nothing else that can be chosen”.

Moreover, evaluating both evidence and reason in one argument distracted students’ attention from the SA element being assessed since their evaluation of *reason* was affected by their evaluation of *evidence*. Items 3 and 5 in Scenario 2 (see Appendix 18) asked students to evaluate Jane’s evidence (which is inappropriate) and reason (which is based on the inappropriate evidence, although it is reasonable for the specific evidence) respectively, then student’s evaluation of *reason* was affected by their evaluation of *evidence*. For example, ST1 said, “the evidence is not complete for she ignored other evidence, so the reason is neither

reasonable nor comprehensive” and ST4 said, “it is not the only evidence, so I chose D (none of above) for her reason”. Additionally, including different types of SAC elements in one task/scenario confused students by requiring them to switch between different competences. Students sometimes forgot what they were asked to do, such as SF1 who said when dealing with an E-SA item that “my mind was still staying on identifying SA, I need to familiarize myself from identify to appraise”.

Thirdly, it seemed difficult for students to clarify what the test was asking them to do, and most students (N=6/8) had little awareness of the aim of the test, although the test title (i.e., SAC test) and the assessment scaffold were shown to them. For instance, ST1 said “(I think it is assessing) the understanding about what Bob and Jane say and the scenario”, and ST2 said, “it is not very different from the test before...but I need to read the dialogue and analyze the information carefully”, although some students mentioned that it was assessing their “thinking ability”.

Additionally, some students misunderstood or forgot about the problem needed to be argued in each task when they were responding to items. For instance, the first task (see Appendix 18) was to discuss how water would influence frictional force, but some students mistook it as to discuss why cars move slower on rainy days. For example, after ST1 thought aloud on task 1, I asked her “what do you think their claims are?”, she said that “the topic they keep discussing is why the car slowed down in rainy days”. In the follow-up interview, I asked SF1 “Did you realize that they are discussing a specific question in each task?”, she said that “I realized it a bit at first, but then forgot about it when I kept trying to answer questions”. Moreover, dealing with too many items in one task seemed to have distracted them, such as ST2 said, “I paid too much attention on identifying from their dialogue, and wasn’t aware what needs to be discussed”.

However, when they were reminded about the aim of the test or each task, they would realize what it was assessing. There were three possible reasons for this phenomenon. The first reason was that students have always been taking the traditional form of tests that assess their content knowledge mostly, and they have never been taught explicitly about argumentation in the school context, so they were not familiar with what exactly was being assessed by the test. For example, ST4 said, “we never saw this kind of test before” and ST3 said that “the test we used to take does not need to answer why...we usually use formulas to answer question”. The second reason might be that they were nervous when doing the test. As ST2 mentioned “I felt nervous

and focusing on finishing it, so I did not pay attention to what the test is about”. The last reason was that the test was not designed well to elicit their awareness of the aim of the test.

Fourthly, all participants to some extent mentioned the terms used in the test. Teachers who participated in the interview worried whether students can understand the terms used for SA elements such as *Evidence* and *Claim*. For example, S3 said, “students need to learn what is reason, evidence and rebuttal based on the test, which is very difficult for them...students would wonder what is relevant and comprehensive, and it is difficult as well”. The students seemed to have various understandings about SA-related terms such as relevant, sufficient, and reasonable etc., which were mainly in E-SA items. There are two aspects related to students’ understanding of terms: what these terms mean in general, and under specific context. Students explained the semantic meaning of these terms quite well when they were asked about the general meaning. For example, SF3 said, “relevant is that (evidence) is related to what they are discussing...reasonable means conforming to facts and logical, comprehensive means containing every aspect”, and SF1 said, “comprehensive is to consider every aspect...reasonable is conforming to logic and being acceptable”.

However, when considering these criteria in specific contexts, students had different understandings and some students even contradicted themselves. As reflected from one (SF1) of the participants’ performance: she explained why she didn’t choose ‘reason is comprehensive’ in Task 3 (see Appendix 18) as “it is only reasonable for this item but cannot be applied to other similar items”; and she explained why she chose ‘none of above’ for the Er item in Task 1 as “it does not consider other information so it is not comprehensive, and it is unreasonable because it does not consider other information”. It can be identified that the student didn’t have a clear understanding about these criteria in SA and didn’t connect the criteria to the specific argumentation context.

Fifthly, as was previously revealed in pilot I, the requirement of content knowledge for each item still affected student’s performance. It is difficult for students to engage in argumentation if they do not know the content knowledge, and less information given in the task means that students need to recall more knowledge they have learned, resulting in assessing more about knowledge proficiency. For example, students showed little argumentation process in Task 5 (see Appendix 18) as this problem was more difficult for them, and little information was provided. As indicated by ST4 “there is no information in the task can support my claim...I don’t know, just based on my intuition”, and by SF3 that “I just guessed...by intuition”.



Lastly, the storyline didn't seem to help students, and it may have added extra burden to students for it provided information useless for engaging in the test, as indicated by the students. Overall, the problems of the test revealed by this pilot were:

- 1) test was too long;
- 2) items were dependent;
- 3) test/task aims were not made explicit to the students;
- 4) SA-related terms were not clarified;
- 5) items didn't provide enough useful information;
- 6) the test design distracted the students.

#### **5.4.2 Test modification**

In a different way from the transformation from Test version I to Test version II, where the key word was 'enrich and expand', the design of Test version III was described as 'abridge and focus' to further elicit students' SAC and prepare the assessment for a larger scale. Based on the findings illustrated in the previous section, the following strategies were then used: **shortening test length, making the problem to be argued explicit, changing scenario arrangement, resolving item dependence, clarifying SA-related terms, and providing more information.**

Firstly, the story line in the test was deleted, instead the problem that needs to be argued was provided explicitly as the title of each scenario to help students engage in the test (see Figure 5.7, Figure 5.8, Figure 5.9). Another benefit of making the problem to be argued explicit was to reduce the possibility that students may confound SA elements. This was because the function of a statement may change in different contexts, for instance, reason may become claim when the problem to be argued changes. Also making the problem to be argued explicit makes SA elements that need to be proposed clear and consistent within a task.

To solve the interdependence of items under the same scenario and the students' confusion caused by engaging in different cognitive processes (i.e., different SAC component is needed), each task was designed as assessing mainly one SAC component, and students were required to evaluate only one element in each argument. To eliminate the possible interference of SA elements in the same argument, the remaining elements that do not need to be evaluated were kept appropriate. To make the assessment more supportive and eliminate the threat of unfamiliarity to assessment validity, the definition of SA was provided as assessment scaffold

and a lead-in I-SA task was presented as the first task. Additionally, according to the findings in pilot I and pilot II, several inevitable task characteristics could influence students' engagement in the items, namely the required content knowledge, the provided information, and the familiarity to a topic. So, these factors were deliberately considered when designing and modifying the test. Specifically, the Tasks in Test version III were arranged in the order of I-SA, E-SA and P-SA and simple tasks therefore were followed by more complex tasks (i.e., requires more content knowledge or provides more information to compare or with unfamiliar topic) within each category. Additionally, a title was provided before the first task of each component, allowing students to engage in the assessment progressively (see Figure 5.7).

The test consisted of less items, and much less I-SA items were included since almost all the students could get I-SA items correct in the previous pilots. A social science issue was included to explore the students' performance on items that do not necessarily need content knowledge. More pieces of information, both relevant and irrelevant, were listed *explicitly* in each task to enable students to compare and use evidence.

**Explanation of the four elements of scientific argument:**

**Claim:** an opinion or conclusion on an issue.

**Evidence:** data or facts used to support a claim.

**Reason:** an explanation of the link between evidence and an opinion, i.e., why a certain piece of evidence supports a certain claim.

**Rebuttal:** questioning and weakening the arguments of others.

**I Identification of argument elements—This is to see whether you can identify the four elements of argument in a piece of given argumentation**

**Problem 1: Which has greater inertia, ball A or ball B?**

	Velocity	Mass
Ball A	5m/s	5kg
Ball B	2m/s	10kg

1) Inertia is an inherent property of objects, which is only related to mass. The greater the mass, the greater the inertia.

2) The mass of ball B is 10kg, which is larger than that of ball A.

3) The ball with high velocity is less easy to stop, and it is not necessarily only related to mass.

4) Ball B has greater inertia.

**Please select from 1-4 to fill in the brackets below:**

Claim is (    ) ; Evidence is (    ) ; Reason is (    ) ; Rebuttal is (    )

*Figure 5.7 Scaffold and I-SA task 1 (Test version III)*

To make the options of E-SA items more understandable for students, the options were elaborated in a more specific and closer to context way (see Figure 5.8). One more option was added to Erb items (e.g., Jane provides her own claim) to capture a situation that often occurs

for students in that they do not engage in other’s argument (Romine et al., 2020; Chen et al., 2019).

**II. Evaluate the elements of argumentation (single choice) - judge whether the elements of argumentation conform to the given indicators**

**Problem 3: Does water increase or decrease friction?**

**“I think water increases friction,”** says Bob. **“When counting money, it’s easier to count with your fingers dipped in water.”**

Jane said: **“You are too one-sided, and the tires tend to slip when the road is wet.”**



1. Bob believes that the single-underlined text is his evidence, which of the following do you think his evidence fits into? ( ) (one or more choice)
  - A. Bob’s evidence is relevant to his claim
  - B. Bob’s evidence is sufficient to prove that his claim is right
  - C. None of the above
2. Jane thinks that the double-underlined text is her rebuttal to Bob. Which of the following do you think her rebuttal fits into? ( ) (one or more choice)
  - A. Jane points out Bob’s deficiency
  - B. Jane proves Bob’s deficiency with appropriate evidence
  - C. Jane provides her own claim and explains about it
  - D. All of what Jane says is right
  - E. None of the above

*Figure 5.8 E-SA task 1 (Test version III)*

**III Production of argument—This is to see whether you can formulate your own argument when facing scientific issues**

**Problem 5: Should it be forbidden to play toy helicopters on the train?**



**Some facts about motion, trains, and toy helicopters:**

- a. The helicopter remote control has three function buttons: ascend, descend, and keep hovering.
- b. Objects that are not subject to external forces tend to maintain their original state of motion.
- c. There will be acceleration when the train starts and during its running.
- d. The train has a maximum speed limit.

**Bob: “I think it should be banned. It is very dangerous to play helicopters on the train since it will bump into people or the carriage door.”**

**Jane: “I don’t think it should be banned. The luggage on the train does not slide, so the helicopter will float stably after it rises, and there is no danger.”**

1. What is Bob’s claim?
  - A. He provides no claim B. His claim is \_\_\_\_\_
2. Which fact(s) from a-d could be used as evidence to support what Bob says?
3. Why do the fact(s) you choose support Bob’s claim?
4. How would you rebut Jane?

*Figure 5.9 P-SA task 1 (Test version III)*

**5.5 Test version IV- Finishing touches**

Pilot III didn’t reveal any further problems in the test, and students showed engagement in SA

with clear awareness of what they needed to argue about, and they were not confused about the test design. So, the test was administered in a large-scale pilot. This section mainly introduces the problems revealed from pilot IV.

### **5.5.1 Main findings from pilot IV**

The main findings obtained from the IRT analysis of the test scores was the poor performance of E-SA items (see section 6.2.1.5). As mentioned in section 5.3.2.2, one or more options in E-SA items could be selected since an SA element may meet several criteria represented by the options. The scoring of E-SA items was thus time consuming since it needed to be recognized that whether the response include

- 1) options that are all incorrect;
- 2) both correct and incorrect options;
- 3) correct options but not complete; or
- 4) completely correct options.

IRT analysis showed that E-SA items didn't seem to differentiate students who got different scores well. The follow-up interview with students revealed that two students didn't recognize the relationships between the Ee and Er item options. Such as F10 who said, "I misunderstood the (Ee) item mainly because I didn't recognize that the options are correlated, and I thought they are describing from different aspects". Furthermore, the Erb items involved more options and aimed to capture different aspects of an inherently more complex rebuttal. However, despite their demonstrated understanding of each option, overall, participants seemed to dive headfirst into the details without realizing the most important features of a rebuttal (i.e., engaging in others' arguments and weakening/being persuaded). For instance, when the participants were talking about Ee items, they evaluated evidence by connecting it with the claim such as "the evidence indeed supports the claim, but it is not enough on its own because there are other situations that can falsify the claim, so the claim should be argued to be correct by combining with more evidence" (F7). However, they tended to analyze each option in Erb items without integrating the rebuttal in the whole argumentation context.

Some language problem and some provided information in Task 9 were revealed as confusing for students. Such as two students showed in the interview that they had limited understanding about 'Environmentally friendly materials'.

### 5.5.2 Test modification

Based on the findings, the key words of the modification were **clarifying SA-related terms** and **changing item format**. Options in the E-SA items were modified to try to eliminate students' misunderstanding about these options and offer the students a clear picture of 'What is the meaning of each option and therefore what is important for each SA element'. Specifically, the progressive relationship between options in each item was further elaborated and each option was further modified to make it more understandable (see Figure 5.10). Moreover, only one option was correct for each item and the 'none of the above' option was deleted to make it easier to score the items and clear to students. This was thought to be able to improve assessment validity (DiBattista et al., 2014). Corresponding to the modification of the E-SA criteria, scoring rubrics for P-SA items were modified to make it clearer the most important point in each scoring category (see Appendix 22).

The criteria for *Evaluating rebuttal* (Erb) items were changed to whether the rebuttal 1) **attended to** and 2) **weakens** the opposite argument. This was due to the intention to evaluate the core characteristics of rebuttal and the complexity of designing items for using too many criteria. Previous studies also emphasized the importance of attending to and understanding each other's argument (Romine et al., 2020; Berland & Reiser, 2011). In addition, items with poor performance in terms of discriminating between students (i.e., extremely easy) were deleted, and the language of each item were further polished. Example of an E-SA task is shown below. A final test specification can be found in Appendix 20.

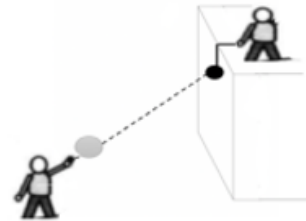
## II. Evaluate the argument (single choice) - judge whether the elements of argument meet the given criteria

### Problem 3: Where to aim to hit the black ball?

Three students are participating in a throwing competition, and the player who hits more black balls with grey balls wins. When the participants stand on the ground to throw the grey ball, one person releases the black ball from the high platform at the same time.

**Some facts about the two balls:**

- Black balls are heavier than the grey ball
- Black balls are released without initial velocity
- Air resistance has little effect on the balls' motion
- If the grey ball is not subject to gravity, it will move in a uniform straight line in the direction of throwing (dotted line)
- Gravity causes the vertical displacement of the grey ball to decrease by  $gt^2/2$  compared to when without gravity



**Bob said:** “Aiming below the black ball is easier to hit. Because the black ball has larger mass.”

1. Bob believes that the single-underlined text is his evidence, which of the following do you think his evidence fits into? ( )

- The evidence is irrelevant with his claim and cannot support the claim
- The evidence is relevant to his claim, showing that his claim may be right, but not sufficient
- There is sufficient evidence to establish that his claim is right

**Jane said:** “I also think we should aim below the black ball. According to facts b and c, I have the feeling that the black ball will fall faster, and the grey ball can only hit it by aiming down.”

2. What is Jane's evidence? ( )

- She provides no evidence
  - Her evidence is: (please mark it with single underline)
3. Jane thinks the text marked by dashed line is her reason, which of the following do you think her evidence fits into? ( )
- There is no connection between the reason and her evidence
  - The reasons explain her evidence correctly, but cannot suggest that the evidence proves her claim being right
  - The reason thoroughly illustrates the connection between her evidence and claim

**Li said:** “I disagree with you two. I think we should aim directly at the black ball. Facts b, c, d, and e can support my claim.”

4. Li thinks that the double-underlined text is his rebuttal. Which of the following do you think his rebuttal fits into? ( )

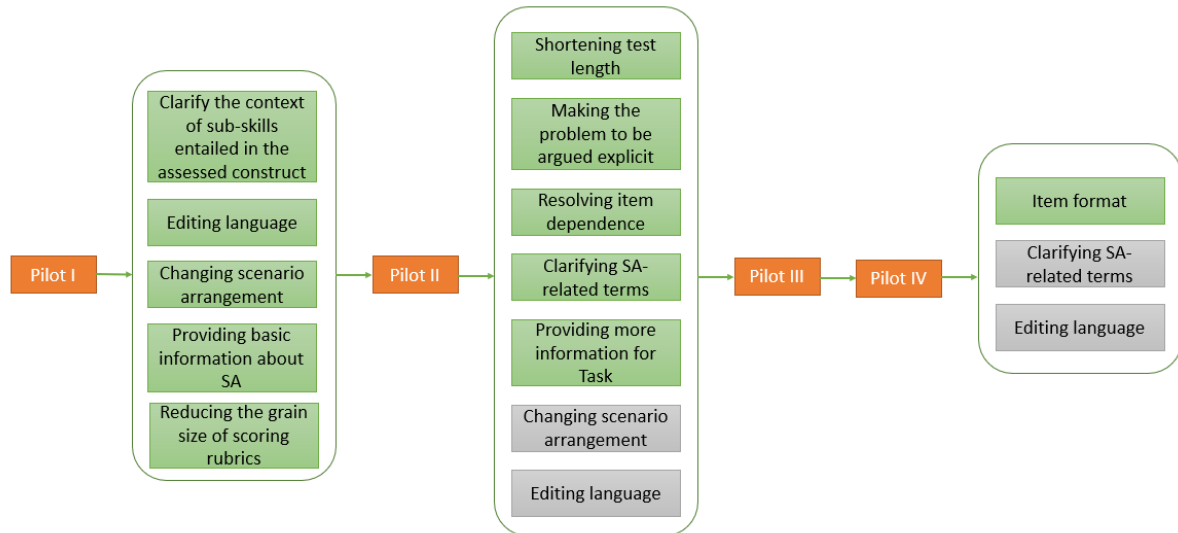
- Li provides his claim without paying attention to Bob and Jane's argument
- Li analyses Bob and Jane's argument without weakening their argument
- Li weakens Bob and Jane's argument

Figure 5.10 E-SA task 2 (Test version IV)

## Chapter summary

This chapter explicated how the assessment instrument was developed in an iterative process by highlighting the empirical evidence obtained from each pilot study that contributed to the improvement of the assessment, which answered RQ 1. 11 factors, some of which appeared in more than one pilot (grey box), were found to influence the assessment quality and therefore corresponding strategies were adopted (see Figure 5.11). It turned out that Wilson's (2004)

four building blocks and the iterative process it emphasizes, as a general guidance, indeed helped the modification of a SAC assessment instrument. However, items design is a much more tedious work that need more specified guidelines for a specific area such as SA.



*Figure 5.11 Strategies employed to improve the SAC assessment*

This chapter made the process of developing an SAC assessment transparent, as well as the product generated by each modification. The documentation in this chapter serves as evidence that can be used to support the micro-validation of the SAC assessment (see section 3.5.2), which will be elaborated in the next chapter. Overall, this chapter explained how a SAC assessment was developed to illustrate how this study understood SAC based on literature review and implementing pilot studies. The next chapter will use empirical evidence to evaluate this understanding, thereby justifying/modifying/expanding it.

## **Chapter 6. Validating the SAC Assessment and Understanding the SAC Construct**

### **Introduction**

This chapter aims to answer Research Questions 2 and 3, that is, to justify the interpretation of the assessment results, so as to legitimize the expanded understanding of SA based on the assessment. The purpose of the assessment was to explore the assessment of SAC from three components (i.e., I-SA, E-SA, and P-SA). Thus, the overarching claim of the assessment is that *‘It is possible to measure SAC from the three components and by using the SAC test results.’* As mentioned in section 3.5, this study probes the investigation of validity by formulating validity arguments from both a micro and macro perspective. Thus, section 6.1 will present the ‘interpretation/use argument (IUA)’ formed by a network of claims that come from a macro or micro perspective to illustrate how the network can support the overarching assessment claim if all the claims in it are supported. Sections 6.2 and 6.3 will construct the macro and micro validity arguments by evaluating the IUA using empirical evidence (obtained from pilot studies and the main study) and logical analysis. By doing so, the two sections inform the extent to which the process and product of developing and administering the instrument showed validity in supporting the assessment of SAC, namely to what extent the IUA is supported.

Responding to the weaknesses embodied in the validity argument, section 6.4 will further check the SAC test items to generate a final data set from the assessment results that represents the students’ SAC. Section 6.5 will address how the assessment results inform a learning progression of SAC by analyzing and discussing the empirical evidence from this study and previous studies. By doing so, the interpretation of SAC test scores can also be further specified.

### **6.1 IUA of the assessment**

As discussed in section 3.5, the ‘interpretation/use argument (IUA)’ is an argument consisting of a network of claims that illustrates the supposed interpretation and use of an assessment. An IUA may include either the interpretation or the use of an assessment or both, and the claims in it are often different based on the different purposes of the assessment (Kane, 2013). For this study, claims in the IUA are proposed from a macro/micro perspective, the micro-claim is that *“The assessment procedure is conducted effectively and elicits participants’ SAC”*. Unlike the micro-validation, the macro-validation focuses directly on the overarching assessment claim (Newton, 2016), which is *“It is possible to measure SAC from the three components and by using the SAC test results”*. Drawing upon previous frameworks on validation argument



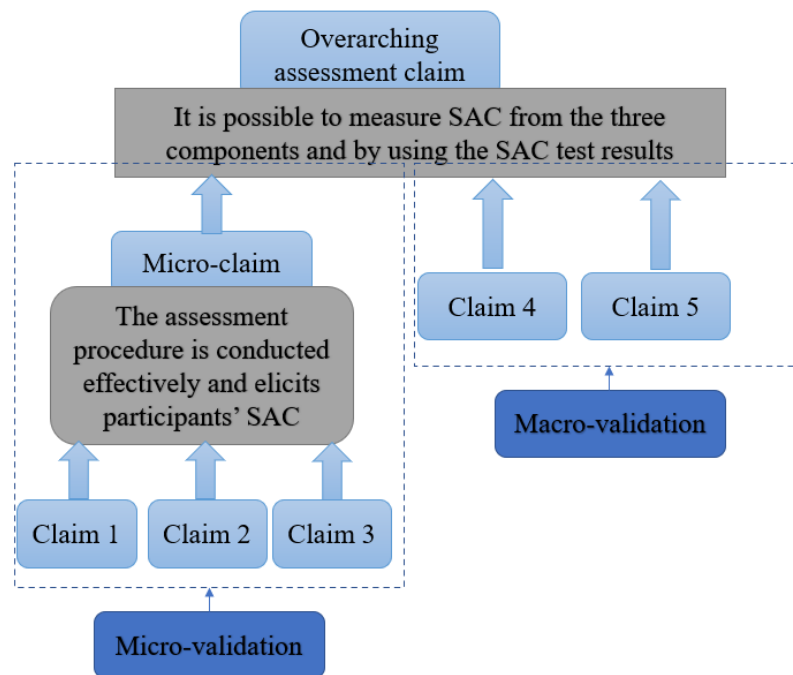
(Ferrara & Lai, 2015; Shaw & Crisp, 2012), three claims work together to support the micro-claim in this study:

- 1) The instrument development procedure produces items that elicit SAC,
- 2) The test administration follows the prescribed procedure,
- 3) The scoring processes are consistent and accurate for all examinees.

Two claims are proposed to support the assessment claim from a macro perspective:

- 4) The internal structure of the construct is represented accurately in the assessment,
- 5) There is no negative impact on the participants by implementing the assessment.

Their relationship is shown in Figure 6.1 below.



*Figure 6.1 Micro and Macro validation for the SAC assessment*

In which Claim 1 has its own sub-claims to be justified. The complete list of claims is presented in Table 6.1 below.

*Table 6.1 Claims in the IUA of the SAC assessment*

<b>Claim</b>	<b>Illustration</b>
Claim 1	The instrument development procedure produces items that elicit SAC
1-1	The item writer understands the assessment target and how to write items to elicit SAC
1-2	The instrument development procedure and tools support the item writer in focusing on SAC
1-3	Items align well with the construct map and Items design
1-4	Items elicit SAC
1-5	The instrument developed by the iterative procedure is ready for the main administration
Claim 2	The test administration follows the prescribed procedure
Claim 3	The scoring process is consistent and accurate for all examinees
Claim 4	The internal structure of the construct is represented accurately in the assessment
Claim 5	There is no negative impact on the participants by implementing the assessment.

Figure 6.2 illustrates how these claims relate to each other to support the overarching assessment claim thus form the IUA for the validation of the assessment. Claims 1 to 3 are focused on the micro-validation and claims 4 and 5 are validating the assessment from a macro perspective. Given that the micro perspective is concerned about the process of developing an assessment, claims 1 to 3 each focuses on one stage in the assessment development process before the administration of the assessment. In contrast, the macro-validation focuses on the outcome of administering the assessment.

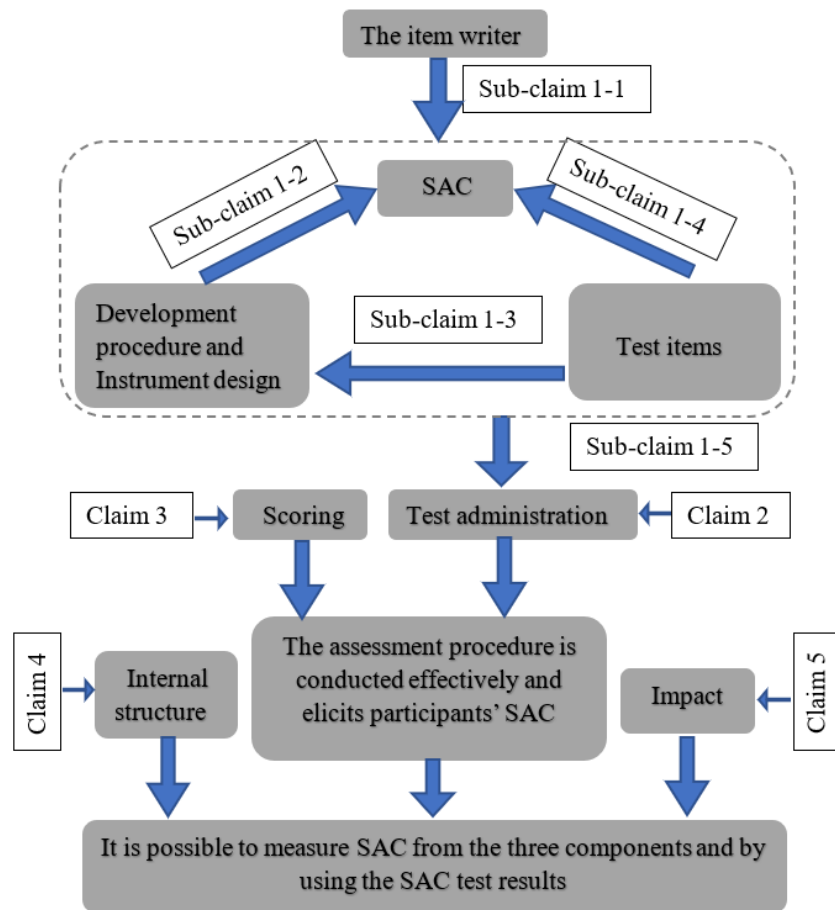


Figure 6.2 The claims network of the IUA for the SAC assessment

Claim 1 is focused on the development of assessment items, in which the item writer (researcher), the assessment target (SAC), the method that guides assessment development, and the designed items/test interact with each other. A proficient understanding toward assessment and the assessment target is a vital and an initial step for the endeavor of developing an assessment (Wilson, 2004), thus the item writer should understand and articulate SAC and know how to develop an assessment (sub-claim 1-1). Based on the conceptualization of SAC and the knowledge on assessment, the item writer should make the procedure of assessment development and the design of the assessment instrument appropriate for SAC (sub-claim 1-2). Then, the items should be designed following the procedure and should be consistent with the prescribed construct map and items design (sub-claim 1-3), otherwise the prescribed procedure and design are meaningless. Even if the items are designed to be consistent with the prescribed specification following the prescribed procedure, it does not mean that the items can assess the targeted competence. So, it is necessary to justify that each item can actually assess what it is supposed to assess (sub-claim 1-4). The dashed box therefore represents the items

development process. Finally, the instrument as a whole should be appropriately prepared so as to be assessing SAC and ready for large-scale administration (sub-claim 1-5). If claim 1 is justified, the test paper should be able to elicit the students' SAC, claim 1 thus needs to be justified first before going on to claim 2.

Administration can be misconducted to influence the results of assessment and thereby its validity (AERA, 2018), so the test should be administered in an appropriate way to minimize potential negative influence on the assessment results (claim 2). If claim 2 is justified, the responses on the test paper should be a trustworthy reflection of the students' performance. After the test paper is collected, it should be scored in an accurate and consistent way across all the participants to ensure that the data truly represents the characteristics of the responses on the test paper (claim 3). If claim 3 is justified, the scores/data that obtained should be an accurate representation of SAC. If all the stages entailed in the above claims are conducted in an appropriate way, then the micro-claim that '*The assessment procedure is conducted effectively and elicits participants' SAC*' should be supported.

To make the assessment claim valid, it is also important to justify the results of administering the test to make sure that the test represents the construct (SAC) accurately (claim 4). Lastly, as discussed in section 3.1, any negative influence on the students brought about by an assessment should be carefully checked for to avoid violating the ultimate aim of promoting teaching and learning (claim 5). This is not the only way to construct an IUA for an assessment, however this study constructs it like this based on the purpose of the SAC assessment and the evidence that can be obtained. The evidence used for the macro-validation is mainly about the psychometric characteristics of the test obtained by conducting a PCM analysis (i.e., using Partial Credit Model), as introduced in section 4.5.3. There could be other evidence that can be used for a macro-validation, such as the relations to other variables (AERA, 2018) and evidence related to uses/consequences as mentioned by Newton (2016). Due to the limitation of time and resources, no evidence on the relations between SAC assessment results and other related variables was obtained. Given that this research is an exploration of a possible way to assess SAC and has not been used officially, no evidence of its use or the consequence of its use was obtained in this study. However, as mentioned in section 3.5.3, the interviews of students' experience of taking the assessment are taken as evidence to shed light on the **possible** impact the assessment may have on students. Taken together, if all the above claims are well supported, both the process and the product of the assessment will have shown validity in supporting the

overarching assessment claim.

The IUA above illustrated why these claims are important and need to be supported to validate the assessment. Section 6.2 and 6.3 will explicate how each of the claims is evaluated by the empirical evidence obtained in this study to inform the extent that the IUA is supported and thereby the extent that the overarching assessment claim is supported.

## **6.2 Validity argument for the micro-validation**

This section aims to validate the assessment from a micro perspective to make the logical chain within the assessment process visible. A validity argument, as mentioned in section 3.5, will be constructed in this section from a micro perspective by evaluating (supporting or falsifying) the IUA critically using evidence from multiple sources and various analysis. The claims related to the micro-validation in the IUA, as shown in Table 6.1, will be evaluated one by one.

### **6.2.1 Claim 1 The instrument development procedure produces items that elicit SAC**

As discussed in section 6.1, the first claim focuses on the development procedure in which the instrument was produced, and five sub-claims are proposed to clarify the relationships between the objects in the instrument development process to form a network to support claim 1. Each of the sub-claims and its arguments are presented below.

#### **6.2.1.1 Claim 1-1 *The item writer understands the assessment target and how to write items to elicit SAC***

Given the construct needed to be assessed, the item writer should possess sufficient understanding of SA, high school Physics content knowledge, knowledge of Chinese context, and educational assessment. Three pieces of evidence are used to evaluate claim 1-1:

- 1) The item writer understood SA and relevant content knowledge,
- 2) The item writer understood assessment,
- 3) The item writer understood the Chinese context of education.

For **Evidence 1**, the literature review presented in section 3.2 drew on various sources and has clarified the understanding of SA from several perspectives. The reviewed literature has certain influence in the field and has been referenced by many studies. Although the illustration of argumentation in this assessment may not be completely comprehensive given it is a broad area, section 3.2 accounted for the importance of the three components that were assessed and

elaborated how they were understood. Also as mentioned in section 1.1, the researcher has a background in Physics education.

For **Evidence 2**, as shown in section 3.3, various studies related to the assessment of SA were critically reviewed and ideas in terms of advancing the assessment of SA were elaborated. These peer-reviewed studies were empirically or logically justified and are worthy of reference. Moreover, several books related to assessment methods, which are systematic and operationally helpful, were reviewed. Wilson's (2004) approach was selected as the principal approach to guide the assessment of this study due to its advantages at measuring cognitive skills and elaborating each step of an assessment with supportive practice material (see section 5.1). Methods to validate an assessment were compared therefore the appropriate methods to use in this study were explicated in section 3.5. Despite referring to these theoretical resources, experience and practice is importance for an item writer. I don't have much experience as an item writer, but this study adopted an iterative procedure to obtain empirical evidence to improve the assessment at each iteration, which makes up for the potential disadvantage to a certain extent. For **Evidence 3**, Chapter 2 elaborated the Chinese context with a focus on its examination culture.

The argument above shows that the potential rebuttals were anticipated, and efforts were made to reduce its influence by appropriate research design and procedure. Thus, claim 1-1 seems well supported in that *'The item writer understood the assessment target and how to write items to elicit SAC'*.

#### **6.2.1.2 Claim 1-2 The instrument development procedure and tools support the item writer in focusing on SAC**

Given claim 1-1 that the researcher had suitable knowledge to construct an SAC assessment has been justified, claim 1-2 emphasizes that the instrument development procedure and relevant tools (i.e., prescriptions of the Items design) should be appropriately designed to guide the assessment of SAC in the right direction. Two pieces of evidence are used to check the claim:

- 1) The development of the instrument adopted an iterative process which included four pilot studies and a main study,
- 2) The nature and characteristics of each assessment version were specified.

**For Evidence 1**, as illustrated in sections 5.2 to 5.5 , each pilot had its own aim to be achieved

to work towards designing a high-quality instrument and was carefully conducted following Wilson's (2004) approach for assessment. The data in each pilot study were carefully analyzed, and problems that influenced students' interaction with the test and the quality of the instrument were documented and resolved so that the instrument could be improved and thus used in the next pilot study. According to the students' interview data in the third pilot, students got much less confused about the test and provided more positive feedback when interacting with the revised test. No problems with the instrument were revealed in the third pilot, although the instrument quality may have been even higher if more pilot studies were conducted, and more participants were invited in each study. All in all, as an exploration of assessing an under-operationalized construct, the iterative process helped recognize and address potential problems in the assessment.

In terms of **Evidence 2**, the characteristics of the instrument were specified explicitly to make the design of each version of the instrument systematic. As shown in sections 5.2 to 5.5, the characteristics of each version of the instrument were decided upon based on the empirical data from the previous pilot, in which the clarification of the Construct map and the modification of the Outcome space all contributed to the design of items. Thus, the design of each version of the instrument supported designing/modifying test items well.

Overall, claim 1-2 seems well supported by the evidence in that '*Instrument development procedure and tools supported the item writer in focusing on SAC*'. Although there may be better or more creative ways of designing the items, it is reasonable to argue that this is not a factor with the potential to undermine the plausibility of the assessment.

### ***6.2.1.3 Claim 1-3 Items align well with the construct map and Items design***

The previous argument justified that the pre-designed development procedure and Items design were appropriate to support the assessment of SAC. Claim 1-3 concerns whether the items produced could work consistently with the items design (i.e., description of the construct of SAC and the characteristics of items) across each pilot, making each pilot worth conducting towards the preparation of the main administration. Sections 5.2 to 5.5 have elaborated the items design of each assessment version and how test items/tasks were modified in accordance with the items design, three pieces of empirical evidence are used to evaluate this claim:

- 1) Findings from the instrument review,
- 2) Findings from the students' follow-up interview,

### 3) Findings from the IRT analysis of test scores.

For the instrument review (**Evidence 1**), teachers who were experts in Physics education and have the teaching/research experience in Physics education were invited to review the instrument and the items design. All the 10 teachers agreed that the items were correctly aimed at the sub-skills they were supposed to assess, for example T10 said “I am afraid that I cannot provide more suggestions on how to conceptualize SAC because it looks reasonable to me...when I looked at these items, I indeed think they correspond to your design quite well.” Only teachers who were interested in SA and who were good at critical thinking and creative teaching were invited to assist in the study, but they were not experts in SA. This might impede the further improvement of the assessment but should not undermine its validity.

For the student follow up interviews (**Evidence 2**) these indicated that each SAC component and SAC element was differentiated and assessed as expected. As mentioned in section 3.5.3, test-taker’s experience and perceptions of taking the test can assist in discovering whether the test worked as expected. As described in section 4.4.2, students participated in the follow-up interview within 7 days of taking the test when they still had a reliable impression of it. Participants were not guided by the interviewer to the expected answer, and the interviewer created an equal and relaxed atmosphere during the interview deliberately. Thus, what students said in the interview should be reliable. Students (13 out of 18) in the fourth pilot talked about the progressive nature of the assessment, they found the tasks in the assessment were becoming “more and more complicated” (F5). The first few items are the easiest item type which “just ask us to recognize from the given argument” (F2), and Evaluation items are “more difficult” (F4) because they “have to compare different arguments provided in the task and to judge them” (F2) while the Production tasks require them to “have our own claim and to explain it” (F10). For example, F2 said that:

“I felt that these items are arranged in a way that needs us to do more and more thinking...dealing with the items that ask me to identify others’ argument, it is the easiest part since the argument is provided and all I needed to do was to recognize from them; then I need to judge either Bob or Jane’s argument, and probably to use one side’s argument to rebut the other. In the end, I probably need to rebut both sides’ argument and to generate my own claim. I felt myself was arguing especially in the last few tasks where I needed to propose my own claim and to argue about it.”

The finding suggests that the three-component of SAC were realized by participants and their understanding about the difference between the three components confirmed the earlier literature that emphasize the epistemic aspect of SAC (Kuhn et al., 2013; Rapanta et al., 2013).



When they were asked about the difference between items assessing different elements, almost all of them realized that “these items are all about these four elements” (F1). Some of them provided a detailed explanation, for example F4 said:

“When we need to propose our argument, we first need to have our own claim, then we were asked to list our evidence and to analyze the evidence, lastly we need to think from another angle to rebut others.”

The IRT item analysis (**Evidence 3** - see Table 6.2) showed that the estimated difficulty of most items in the fourth pilot were as expected. The fourth pilot data was analyzed using the two-parameter logistic model and generalized partial credit model (2PL + GPCM) and the data fit the selected model according to the SX2 statistics (no statistically flagged items with a  $p < .002$  after Bonferroni correction). Table 6.2 shows each item, its corresponding difficulty, and the targeted SAC element of each item (Evaluation of evidence = Ee; Evaluation of Reason = Er; Identification = I; Identification of claim = Ic; Production of rebuttal = Prb etc.). The more difficult the item, the larger the corresponding parameter. Therefore, most items followed the pattern as expected in that P-SA items are the most complex and I-SA items are the easiest.

*Table 6.2 Item difficulty estimates (fourth pilot)*

<b>Item</b>	<b>SAC element</b>	<b>Estimated difficulty (logit)</b>
I1	I - Identification	-14.4
I94	Er - Evaluation of Reason	-13.7
I2	I - Identification	-5.2
I31	Ee - Evaluation of evidence	-4.9
I51	Ic - Identification of claim	-3.8
I52	Pe - Production of evidence	-2.3
I93	Erb - Evaluation of rebuttal	-2.2
I81	Irb - Identification of rebuttal	-2.0
I82	Er - Evaluation of Reason	-1.2
I43	Er - Evaluation of Reason	-0.9
I41	Ee - Evaluation of evidence	-0.6
I72	Pe - Production of evidence	-0.5
I32	Erb - Evaluation of rebuttal	-0.5
I44	Erb - Evaluation of rebuttal	-0.3
I92	Ee - Evaluation of evidence	-0.3
I61	Pe - Production of evidence	-0.1
I73	Pr - Production of reason	0.1
I53	Pr - Production of reason	0.1
I91	Ir - Identification of reason	0.2
I74	Prb - Production of rebuttal	0.4
I42	Ie - Identification of evidence	0.4
I54	Prb - Production of rebuttal	0.7
I95	Pr - Production of reason	0.7
I84	Prb - Production of rebuttal	0.8
I62	Pr - Production of reason	1.1

I96	Prb - Production of rebuttal	1.4
I63	Prb - Production of rebuttal	1.5

Overall, Evidence 1 and Evidence 2 supported claim 1-3 reasonably with low risk of potential rebuttals. However, in Evidence 3, some items showed unexpected performance (i.e., some E-SA and I-SA items were found to be more difficult than expected while some P-SA items were easier than expected). So, the claim that ‘*Items aligned well with the Items design*’ seems not to be fully supported. It is clear therefore that there is more to item difficulty than purely the type of SAC being tested which one would expect.

However, as mentioned in section 5.2.1, although this study assumed a general pattern in terms of the three components, detailed relationship needs to be informed by empirical evidence. The general difficulty distribution pattern of these items was aligned with the items design and that any unusual items that didn’t fit the pattern were further checked based on students’ interview data before the main study was conducted. So, it seems reasonable to argue that “*Items aligned appropriately with the items design*” before the main administration.

#### **6.2.1.4 Claim 1-4 Items elicit SAC**

The previous argument justified that the designed items were relatively consistent with the construct map and items design, indicating that the assessment was coherent within its own design system. To build a direct connection between the competence and the items, the items should be able to elicit participants’ SAC to rationalize its administration in the main study. Four pieces of evidence are used to evaluate this claim:

- 1) Cronbach’s alpha,
- 2) The results of factor analysis in the fourth pilot,
- 3) Findings from the students’ think aloud interview,
- 4) Findings from the students’ follow-up interview.

**Evidence 1** shows that the Cronbach’s alpha was 0.81 in the fourth pilot, and most items had high item-total correlations (i.e., 0.30 and higher) (see Table 6.3). However, some items had relatively low item-total correlation, indicating that each of these items had a low correlation with the whole test. These items were further checked before the main administration. A value of Cronbach’s alpha higher than 0.70 indicates good internal consistency of the items in the scale, but it does not mean that the scale is unidimensional (Gliem & Gliem, 2003). So, Evidence 2 will show whether the scale embodied by the test is measuring a single construct.

Table 6.3 Item-total correlation

Item	Item-total correlation
I53 Pr - Production of reason	.62
I62 Pr - Production of reason	.64
I73 Pr - Production of reason	.64
I54 Prb - Production of rebuttal	.53
I63 Prb - Production of rebuttal	.49
I72 Pe - Production of evidence	.52
I74 Prb - Production of rebuttal	.52
I95 Pr - Production of reason	.51
I61 Pe - Production of evidence	.44
I81 Irb - Identification of rebuttal	.43
I84 Prb - Production of rebuttal	.42
I96 Prb - Production of rebuttal	.43
I32 Erb - Evaluation of rebuttal	.33
I42 Ie - Identification of evidence	.33
I52 Pe - Production of evidence	.32
I91 Ir - Identification of reason	.31
I93 Erb - Evaluation of rebuttal	.32
I2 I - Identification	.24
I41 Ee - Evaluation of evidence	.19
I43 Er - Evaluation of Reason	.24
I44 Erb - Evaluation of rebuttal	.24
I51 Ic - Identification of claim	.23
I82 Er - Evaluation of Reason	.21
I1 I - Identification	.14
I31 Ee - Evaluation of evidence	.14
I92 Ee - Evaluation of evidence	.08
I94 Er - Evaluation of Reason	.04

Unidimensional means all the items function in unison and the performance on each item is influenced by the same process. This implies that all the items are assessing one construct. **Evidence 2** is a scree plot from the factor analysis of the students' scores in the fourth pilot, as shown in Figure 6.3. A scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. The point where the slope of the curve is clearly levelling off (the 'elbow') indicates the number of factors that should be generated by the analysis. Figure 6.3 shows a big drop between the first factor and other factors, which can be taken as approximate unidimensional structure of the data (Bond & Fox, 2015). Thus, Evidence 1 and Evidence 2 together indicate that the items in the test had good internal consistency and were assessing a single construct. Nevertheless, the fact that all the items were assessing one construct does not necessarily mean that the construct was SAC. Thus, more evidence that reveals the nature of the construct is needed, which was obtained by investigating test-taker's experience as shown in Evidence 3 and Evidence 4.

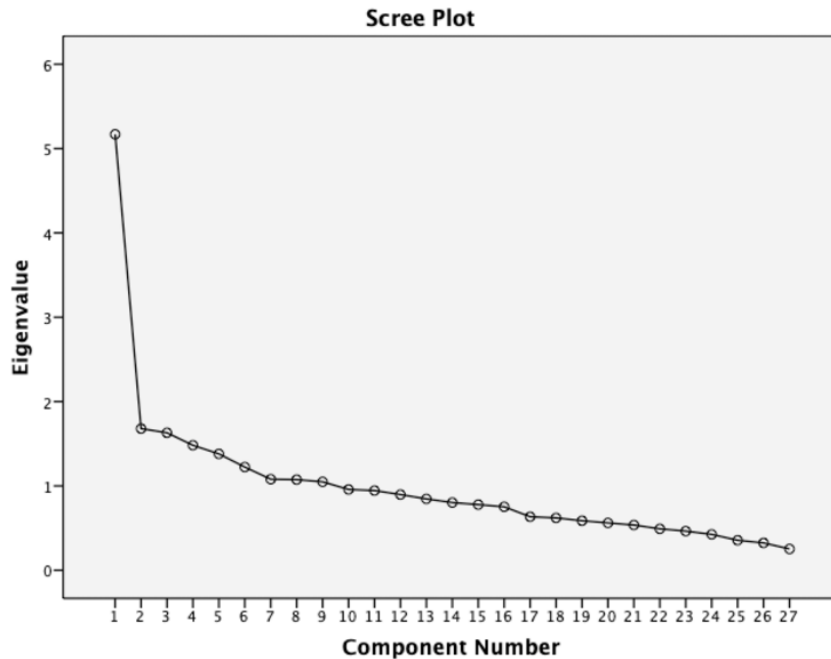


Figure 6.3 Scree plot in fourth pilot

How think aloud data can help identify students' response processes has been discussed in section 4.4.1.2. Students (N=8) who participated in the think aloud in the third pilot showed sufficient engagement in the tasks. In I-SA items, their thinking process indicated that they were using their understanding about the four SAC elements to figure out the functions of the statements in the given argument. As TT1 said in the first I-SA task:

“...this should be his claim since he expresses clearly about his viewpoint on the question...this statement is evidence for he provides an accurate data which is provided in the item stem and is quite explicit...reason should be this one to explain, and this statement challenges other statements, so should be rebuttal”.

Similarly, in E-SA items, all the students were evaluating the given arguments and analyzing them according to the item options to some extent. As TT2 said when thinking aloud item 3.1 (see Appendix 19):

“Bob's evidence is true, and it can indeed support his claim that water can increase friction, but he only provides one evidence. So, his evidence can support him but cannot assure that his claim is right...”.

Students were analyzing the questions and the provided information to propose their own claim and were weighing up the evidence to justify their claim in the P-SA items. Although two out of the 8 students misunderstood the options in some E-SA items, it didn't point to a specific problem with the items.

The obtained data can be taken as evidence because of its fidelity. As described in section 4.4, the process of collecting and analyzing data was conducted in an appropriate and rigorous way. An explanation and demonstration for the participants on how to think aloud using an example was given before asking them to think aloud and they were not prompted about how to respond to the items during the think aloud process. Students were old enough to show the ability to perform think aloud comfortably. So, **Evidence 3** suggests that the interaction between the participants and each item was in accordance with the expectation of designing the item.

**Evidence 4** provides information from a broad perspective in terms of students' experience of taking the whole test. All participants (N=36) in the second, the third and the fourth pilot who took part in the follow-up interviews and the think aloud interviews were asked about their feeling of what they were being assessed on. Given students were not explained in detail what SA is and how the test assessed their SAC, their statements revealed their straightforward experiences of interacting with the test. 31 of them mentioned "logical thinking", 14 of them mentioned "analyze information", 12 students mentioned "argumentation" or "debate", 9 students said that the test assessed their understanding about the four elements of SA, and students also mentioned "language expression ability", "the ability to apply knowledge in real life" and "solve problems in real life". Although students didn't use academic words to describe their perspective, the way they were talking indicated that the items elicited their SAC-relevant skills. When students were further asked explicitly about whether they thought the items were assessing their SAC, most of them provided positive answers with only 3 students saying that they didn't feel anything about SAC. Again, students were not prompted by the interviewer towards any expected answers in the interview. Thus, Evidence 4 suggests that the participants experienced thinking processes related to SA.

Taking all the evidence together, the items elicited participants' relevant skills as expected, and were assessing a unidimensional construct which was very likely to be SAC. There were still some items that showed unexpected performance, such as some multiple-choice items (especially E-SA items) which had relatively lower item-total correlation. These items were reviewed and revised before the main study. So, claim 1-4 seems appropriately supported by the above evidence to be '*Items appropriately elicited SAC*'.

#### **6.2.1.5 Claim 1-5 The instrument developed by the iterative procedure is ready for the main administration**

As justified by the previous argument that the test items elicited students' SAC, this argument

intends to support that the instrument as a product showed acceptable performance to be used in the main study to assess SAC. Evidence about items characteristics and test information resulted from the IRT analysis are used to check this claim.

As mentioned previously the data in the fourth pilot was analysed using the two-parameter logistic model and generalized partial credit model (2PL + GPCM). The coefficient ' $a$ ' in Table 6.4 represents the item discrimination estimate, and ' $b$ '/ ' $b_i$ ' represent item/threshold difficulty estimates. The larger the value of ' $a$ ', the better the item will distinguish between participants who are more proficient at the assessed competence and who are not. For the dichotomous items the larger the coefficient ' $b$ ', the more difficult for the participants to get the item right; and for the items with more possible scores, the larger the coefficient ' $b_i$ ', the more difficult for the participants to transfer from a score of  $i-1$  to  $i$ . Table 6.4 shows that most items have good discrimination (higher than .50), in which P-SA items have higher value of  $a$  and most of the items that had low discrimination value are E-SA items (An & Yung, 2014). In addition, the difficulty estimates of the items cover a wide range of the scale. However, there are reversals of thresholds in five items (i.e., I1-I, I2-I, I31-E, I84-P, I93-E), indicating that it is more difficult for students to get low scores than high scores. For example,  $b_1$  for I2-I is -3.7 while  $b_2$  is -6.7, suggesting that it is more difficult to get a score of 1 than to get a score of 2. So, these items needed to be further modified.

Table 6.4 Item parameters in the fourth pilot

Items	<i>a</i>	<i>b</i>	<i>b</i> <sub>1</sub>	<i>b</i> <sub>2</sub>	<i>b</i> <sub>3</sub>
I41-E	0.3	-0.6			
I42-I	0.7	0.4			
I43-E	0.5	-1.0			
I51-I	0.7	-3.9			
I81-I	1.1	-2.0			
I91-I	0.8	0.2			
I92-E	0.2	-0.3			
I1-I	0.1		-8.4	-20.5	
I2-I	0.3		-3.7	-6.7	
I31-E	0.2		5.1	-14.9	
I52-P	0.7		-5.1	0.4	
I53-P	1.0		-0.8	1.0	
I72-P	1.2		-2.0	0.9	
I73-P	1.6		-0.5	0.8	
I74-P	1.0		-0.1	0.9	
I32-E	0.4		-5.7	-1.0	5.2
I44-E	0.3		-7.6	4.1	2.5
I54-P	1.0		0.1	0.8	1.1
I61-P	1.2		-1.4	1.6	
I62-P	1.6		0.2	1.0	2.0
I63-P	1.0		-0.1	0.6	4.0
I82-E	0.3		-3.0	0.5	
I84-P	0.8		1.4	0.3	
I93-E	0.2		-7.6	7.0	-6.0
I94-E	-0.1		-26.0	-1.6	
I95-P	1.4		-0.6	1.9	
I96-P	1.0		0.5	1.3	2.4

Note: (I: Item; -I: I-SA item; -E: E-SA item; -P: P-SA item. Ordered by item type and Task order)

Visual demonstration of the estimates of each item can be identified from the item characteristic curve (ICC) and category characteristic curve (CCC) in Figure 6.4. In well-performed CCC, each category is centered on a specific ability level and should be peaked, and the categories should be ordered from targeting the low-ability group to the high-ability group as prescribed in the scoring rubrics. In the fourth pilot, the CCC of most items have distinguishable peak for each score category and are dispersed across the ability continuum, indicating most items perform well on discriminating between individuals. Well-performing CCC also indicate that the scoring rubric reasonably categorized different levels of performances.

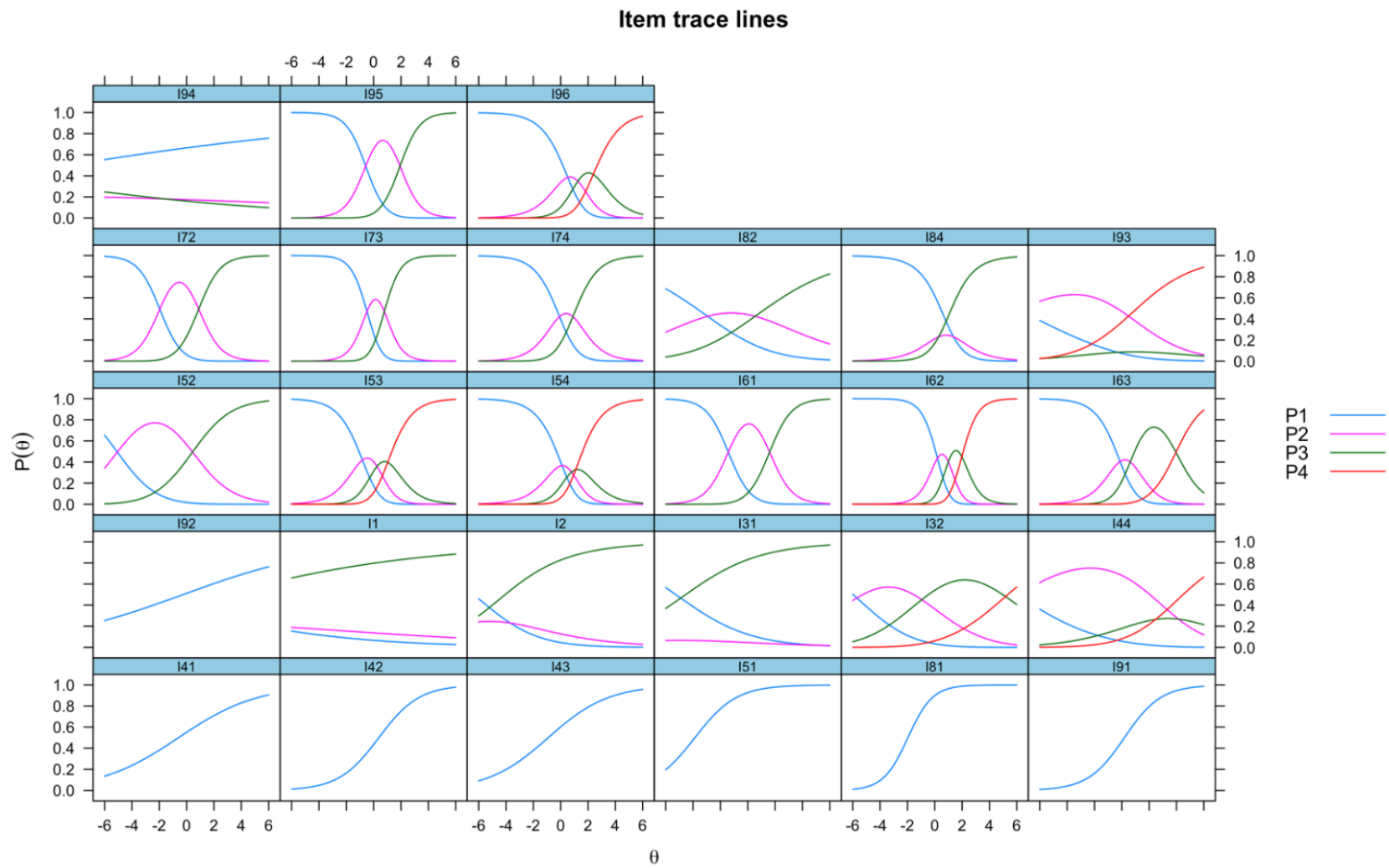
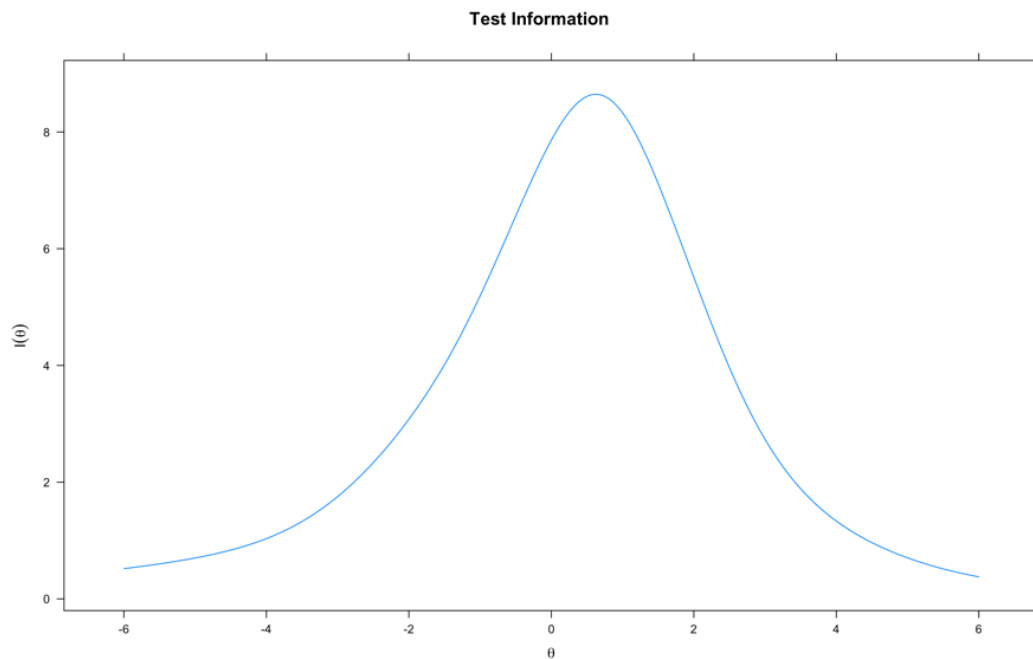


Figure 6.4 ICC/CCC of items in the fourth pilot



Figure 6.5 below is a test information curve (TIC) that shows the range of the scale for which the test provides reliable information. Where the slope of the curve is larger, it indicates that more information is provided for the corresponding participant, and the peak is where most information is provided on the ability continuum. The TIC shows one peak around the middle of the latent trait and has slope between -2 to 4 logits on the continuum. Thus, the test provided more information for students whose abilities ( $\theta$ ) were between -2 to 4 logits, which means that this test targeted appropriately on the ability continuum, and it was more accurate for predicting relatively higher ability students since the peak is to the right of the center. Although item characteristics shown in the previous evidence are not perfect, most items performed acceptably. Items that showed poor performance were modified before the main study. Thus, it is reasonable to claim that *'The instrument developed by the iterative procedure was ready to be administered in the main study'*.



*Figure 6.5 Test information curve (fourth pilot)*

Taking together, claims 1-1 to 1-4 have generally been well supported by the proposed evidence. Claim 1-5 justified the quality of the instrument as a product, but not a final one. Although the product was not perfect, there seemed no damaging threat to the validity of the assessment especially given poor performance items were modified before the main study. As discussed in section 5.5.1, the reason of the poor performance of E-SA items was that some students got confused by the item options in E-SA items, which was revealed in the follow up interviews in the fourth pilot. Thus, it is plausible to claim that *'The assessment development*

*procedure produced items that elicit SAC* before its main administration.

### **6.2.2 Claim 2 The test administration follows the prescribed procedure**

The previous section focused on the instrument development and justified that the items produced in the iterative process elicited students' SAC and were prepared to be used in the main study. The claim in this section emphasizes that the administration procedure of the test should be appropriate and consistent across schools to obtain students' authentic performance. Two pieces of evidence are used to evaluate this claim:

- 1) Administration procedure and
- 2) Students' self-report survey.

For **Evidence 1**, as mentioned in section 4.4.2, the researcher was not allowed to enter schools during the Covid-19 pandemic and teachers helped to administer the test, which brought uncertainty to the administration process. Teachers in each school were briefed about the research and provided with relevant material on the test administration. The researcher emphasized the procedure and precautions of administering the test to teachers. The same test administration procedure was prescribed and given to all the schools. There is no direct evidence from the students' interview to indicate that there were distractions or instances of cheating. However, students' interviews indicate that teachers in different schools followed the administration instructions to different degrees. Some students mentioned that their teacher had been in the classroom supervising them to complete the test while others said that their teacher was not always in the classroom. In addition, some teachers conveyed a more rigorous message to their students and asked them to take the assessment seriously while others seemed insouciant. So, potential threats to the administration procedure are revealed by the evidence, although it does not show any subversive information.

To help evaluate this claim, data from students' self-reported survey are used as **Evidence 2**. As mentioned in section 4.4.1.1, the test paper was followed by a four-question survey and students were encouraged while not forced to answer the questions. If the students did the test seriously and independently, it is very possible that the test was administered appropriately. Thus, this information can be taken as indirect evidence for claim 2.

The result of the survey shows that 1133 (80%) students thought they took the test seriously or very seriously, and 162 (11.5%) students chose the uncertain option. 1290 (91.3%) students

did the test independently while 44 (3%) students were partially independent. Moreover, only 242 (17%) students reported that the time was not sufficient for them to complete the test.

Overall, most students self-reported that they took the test seriously and independently and time was sufficient for them to complete the test. Although there was no direct evidence for the subversion of the administration, the potential rebuttal is uncertain here since any possible additional evidence could make the claim weaker. Thus, it is rational to argue that the administration process could have been better operated and could have been improved. So, the claim of this argument is that “*The test administration generally followed the prescribed procedures*”.

### **6.2.3 Claim 3 The scoring process is consistent and accurate for all examinees**

This claim supports that the scoring process should be monitored rigorously to make sure that the scores given to participants are consistent and accurate. The procedure of scoring and inter-rater reliability is used as evidence to evaluate this claim.

The second rater was a science researcher who was familiar with this study. He participated in the scoring in both the fourth pilot and the main study, and we followed a rigorous procedure to score the test papers. Specifically, we adopted an iterative procedure of scoring in the fourth pilot, we firstly scored a small part of the test paper and then discussed and reached agreement on the specific scoring standard of each item. At this stage, 11 test papers were firstly scored together and there was 40% disagreement between our marking. After an agreement was reached, we marked 94 further test papers together and only 90 items (8%) were scored differently, and the disagreement was thus reduced to an acceptable level. As for the remaining test papers, the researcher scored them twice and 97% of the scores were the same. Since mainly E-SA items were modified after the fourth pilot, the open-ended items and their scoring rubrics were almost the same in the two studies. The above iterative procedure familiarized us with scoring the test, which made us more skilled for the scoring in the main study.

Due to limited time, the second rater only scored a small portion of the test papers in the main study. There were 98 test papers scored by both raters, and Cohen’s *Kappa* of Pe and Prb items were 0.84 ( $p < .001$ ) and 0.76 ( $p < .001$ ) respectively, and that of Pr items was 0.69 ( $p < .001$ ). So far, the inter-rater agreement in both studies were acceptable (McHugh, 2012). The disagreements were further discussed and resolved, and the remaining test papers were scored by the researcher. To maximize the consistency of the scoring, the researcher scored the same

item on all the test papers before moving onto the next item. Moreover, after completing the scoring of test papers of each class, several test papers were randomly selected for a double check. So, it seems reasonable to claim that ‘*The scoring process was consistent and accurate for all examinees.*’

#### **6.2.4 Summary**

Taken together, the procedure of developing the assessment has been designed deliberately and in a rigorous way. Claim 1 has been appropriately supported by the proposed evidence, with some under-performing items modified before the main administration. Claim 3 has also been well supported with a rigorous scoring procedure and acceptable inter-rater reliability. As for a potential rebuttal that not all the items were scored by both raters, it is a plausible one given the limited time and resources and a less detrimental one given that the researcher double checked the scores. However, evidence in claim 2 makes it risky to propose that ‘*The assessment procedure was conducted effectively and elicited participants’ SAC*’. Claim 2 needs more attention because it is rather weak compared to the arguments for other claims, and the potential impact of not being in the field to supervise students is unclear despite the fact that no direct evidence was found to show its impact on the assessment. Thus, the micro-claim is supported properly to be ‘*The assessment procedure was conducted in a moderately appropriate and effective way and elicited participants’ SAC*’.

### **6.3 Validity argument for the macro-validation**

The micro-validation argument in section 6.2 justified that the assessment **process** was conducted in a moderately appropriate and effective way by using evidence collected during this process. This section will validate the assessment from a macro perspective by evaluating claims 4 and 5 that focus on assessment as a **product**.

#### **6.3.1 Claim 4 The internal structure of the construct is represented accurately in the assessment**

The purpose of Claim 4 is to justify that the internal structure of SAC revealed by the assessment is consistent with what was theoretically assumed by the study. Within the framework of Rasch measurement theory (Rasch, 1960), evaluating the quality of measures primarily involves exploring the degree to which item responses reflect the requirements of the Rasch model.

Section 4.5.3.2 has introduced how the data was analyzed. Firstly, Rasch factor analysis results are used to check whether the data meet the uni-dimensionality requirement of the PCM Rasch model. For Rasch factor analysis, a small eigenvalue of the first principal component, usually smaller than 2.0, indicates that the residuals are merely random noise (Raiche, 2005). A larger eigenvalue, on the other hand, implies that there is probably a “second dimension” besides the primary Rasch dimension (Bond & Fox, 2015). Results in this study show that the factor loadings of all the items after extracting the Rasch dimension are within -0.3 to 0.4, and the eigenvalues of the first and second principal component of the residuals are 1.94 and 1.92 respectively, which indicates that the SAC elements formed a unidimensional construct and most of the variability in responses was explained by the PCM model (Bond & Fox, 2015).

As for the potential rebuttal, that the result of uni-dimensionality may be due to the high correlation between items rather than because of SAC being measured by a single construct, an absolute value of  $Q_3$  greater than 0.20 has been suggested as a rule of thumb for flagging an item pair as dependent (Yen, 1993). The results show that four item pairs may be dependent: Pe\_4.1 and Pr\_4.2 with  $Q_3$  equal to 0.36, Pe\_5.1 and Pr\_5.2 with  $Q_3$  equal to 0.30, Pr\_4.2 and Prb\_4.3 with  $Q_3$  equal to 0.28, and Pe\_7.5 and Pr\_7.6 with  $Q_3$  equal to 0.28. Thus, whether a participant can successfully answer one item depends, to some extent, upon that participant’s responses to another item, implying a logical dependency between these items. A possible explanation here could be that as these items share the same scenario, the production of evidence and reason are therefore perhaps unsurprisingly likely to be correlated. However, given overall that few items are correlated with each other, and the coefficient of correlation is not very high, the requirements of the PCM model are not violated.

The next step is to check whether the data fit the model well. Only when the data fit the model, can the analysis results obtained from the model be interpreted accurately within the model framework (Bond & Fox, 2015). Fit statistics and item characteristics are reported in Table 6.5. An acceptable range of MNSQ of within 0.7 to 1.3 has been adopted by many studies, while a lower boundary of 0.6 for human rating items is suggested by Bond and Fox (2015). Almost all the items had acceptable fit statistics except for Erb\_7.2 and Ee\_7.3 which had slightly larger Outfit MNSQ (i.e., 1.39 and 1.43 respectively). These two items need to be further checked. But generally, the Rasch estimates of the items and participants seem appropriate.

Table 6.5 Model data fit and items estimates (main study)

Item	Outfit MNSQ	Infit MNSQ	Location(logit) (SE)	Threshold 1(logit) (SE)	Threshold 2(logit) (SE)	Threshold 3(logit) (SE)
I_1	1.07	1.01	-1.81 (0.07)			
Ee_2.1	1.16	0.98	-2.97 (0.10)			
Erb_2.2	1.26	1.16	-0.07 (0.06)			
Ee_3.1	1.15	1.11	-0.16 (0.06)			
Ie_3.2	0.92	0.95	0.00 (0.06)			
Er_3.3	1.21	1.10	0.80 (0.06)			
Erb_3.4	1.10	1.06	-1.43 (0.06)			
Pe_4.1	0.86	0.86	-0.75	-2.11 (0.08)	0.61 (0.06)	
Pr_4.2	0.72	0.76	0.20	-0.25 (0.07)	0.25 (0.08)	0.59 (0.10)
Prb_4.3	0.76	0.79	0.37	-0.28 (0.06)	0.64 (0.10)	0.76 (0.10)
Pe_5.1	0.82	0.82	0.25	-1.48 (0.06)	1.98 (0.10)	
Pr_5.2	0.67	0.75	1.11	0.38 (0.06)	0.71 (0.09)	2.24 (0.20)
Prb_5.3	0.88	0.90	1.02	-0.02 (0.10)	-0.26 (0.20)	3.34 (0.11)
Pe_6.1	0.95	0.94	-1.33	-2.86 (0.11)	0.20 (0.06)	
Pr_6.2	0.82	0.82	0.66	-0.83 (0.06)	0.66 (0.07)	2.15 (0.17)
Prb_6.3	0.89	0.90	0.41	-1.27 (0.07)	0.64 (0.07)	1.85 (0.14)
Ir_7.1	1.01	1.03	0.07 (0.06)			
Erb_7.2	1.39	1.23	0.54 (0.06)			
Ee_7.3	1.43	1.23	0.55 (0.06)			
Er_7.4	1.26	1.19	-0.31 (0.06)			
Pe_7.5	0.87	0.88	-0.03	-0.84 (0.06)	0.78 (0.07)	
Pr_7.6	0.70	0.83	1.21	0.72 (0.06)	1.31 (0.12)	1.59 (0.21)

From Table 6.5 we can also see that these items and their potential scores cover a difficulty (i.e., threshold estimate) span of around 6 logits from the least difficult (getting Ee\_2.1 correct at -2.97 logits) to that hardest for students to satisfy (achieving a score of 3 on Prb\_5.3 at +3.34 logits), and item difficulties (i.e., locations) range from -2.97 to 1.21, which means that the items cover a large range on the SAC continuum. Moreover, the Wright map (mentioned in section 4.5.3) shown in Figure 6.6 indicates that items matched well with the target population. A Wright map displays the location of item parameters and the distribution of person parameters along the latent trait, which is used to see how well the item difficulty distribution matches the person ability distribution. These graphs are also referred to as person-item maps. The x-axis is the latent trait of the Rasch model (what is measured). On the y-axis is each item. Each item gets a black dot that represents the item location. The white dots represent the location of the thresholds where if an item is dichotomous, it just has the black location dot. Each item's dots are connected by a line. The distribution of the person abilities is at the very top as a histogram - the height of the bar shows how many people are at each ability score. The short black bars beneath the person distribution correspond to the locations of the items.

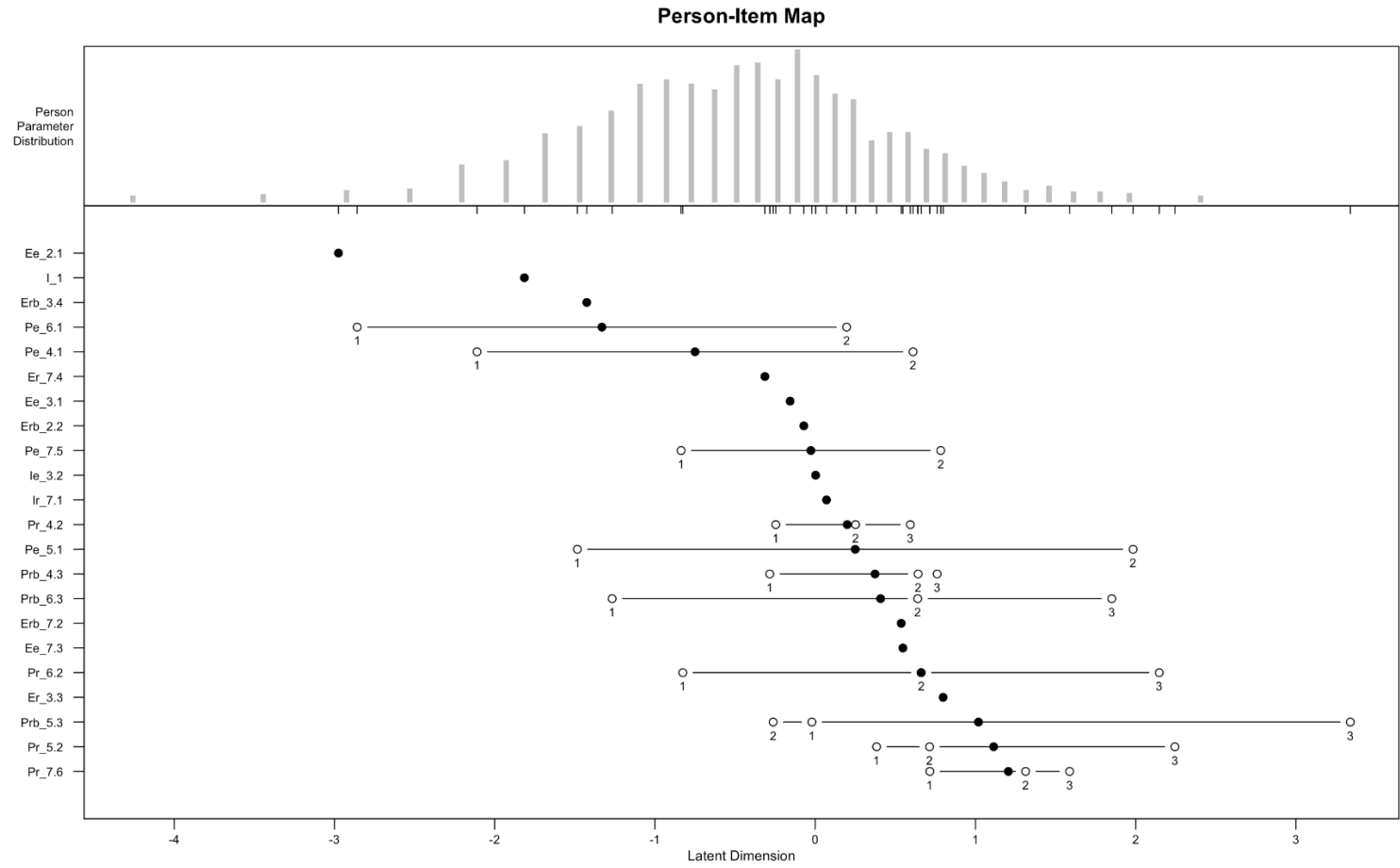
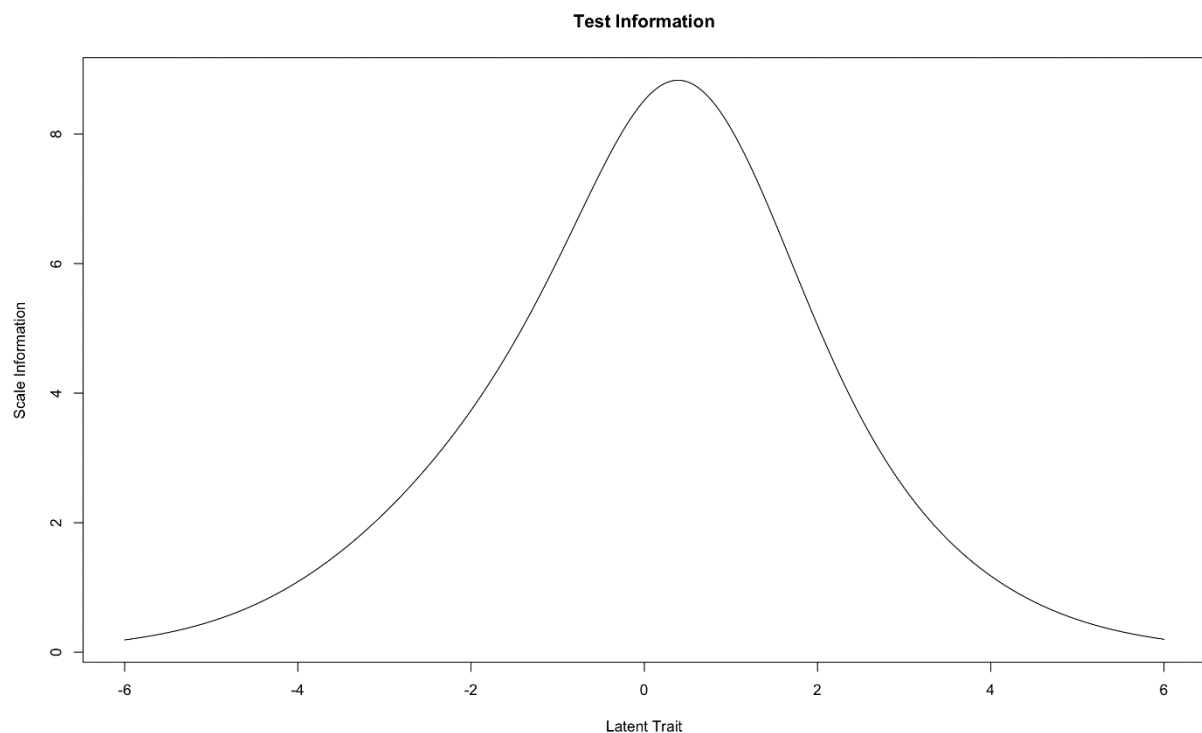


Figure 6.6 Wright map (main study)

Most of the participants in the population had items that had difficulties that matched with their ability. So, the SAC of the group was covered appropriately by these items. However, it seems that most items have a difficulty that ranges from -1 to +1.5 logit, and students' ability estimates have a wider range than where the items are clustered. Thus, the test information curve (TIC) is used to further check whether the test targeted well at the particular group of students. As shown in Figure 6.7, the test is more appropriate to be used to measure students whose ability are between -3 to +3 logits on the continuum. Combining this with the fact that most participant's estimated ability was between -2 to 2 logits as shown in Figure 6.6, the test is appropriately targeted to the sample.



*Figure 6.7 Test information curve (main study)*

In addition, the threshold estimates for most items follow an ordered progression except for Prb\_5.3 for which the required ability to successfully transfer from score 1 to 2 was lower than from 0 to 1 (Table 6.5). Thus, the scoring rubrics for most P-SA items were reasonable and well represented participants with different performance levels of generating SA. As well as the fact that the thresholds in a P-SA item should be ordered appropriately, it would also be better for each category to have a distinguishable peak in the category characteristics curve (CCC) of an item. A peak for each category suggests that the categories are targeted on specific ability level groups, which indicates that the item performs well in discriminating students with different ability levels on generating an argument. Figure 6.8 to Figure 6.11 below show that



most items performed well as each of these categories is actually observed in the collected data and each becomes the most likely observed category as the estimated ability increases along the underlying latent trait. Nevertheless, items like Prb\_4.3 and Prb\_5.3 need to be further examined due to unclear peaks in the category curves.

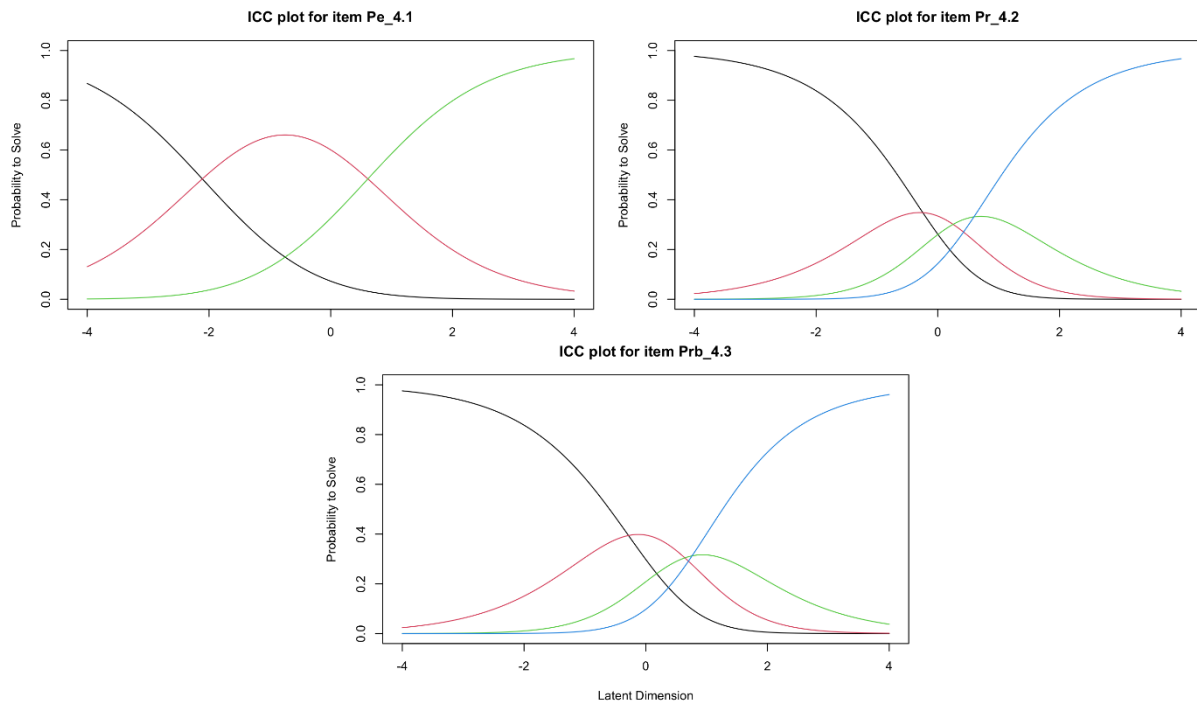


Figure 6.8 CCC plots for task 4

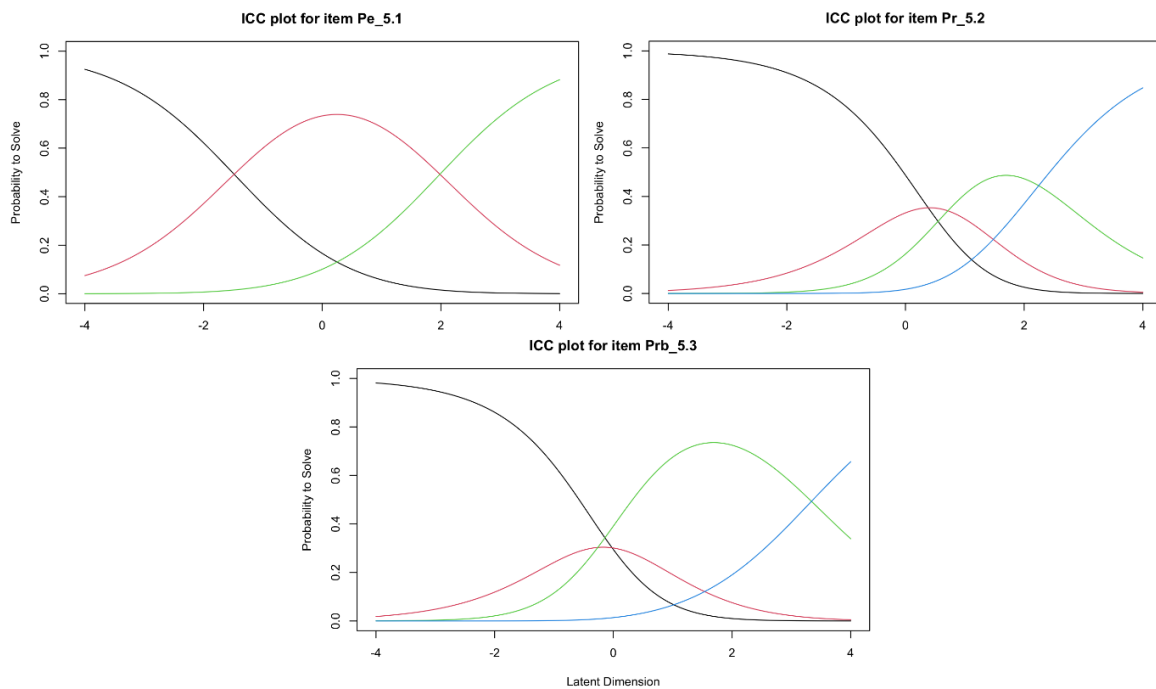


Figure 6.9 CCC plots for task 5

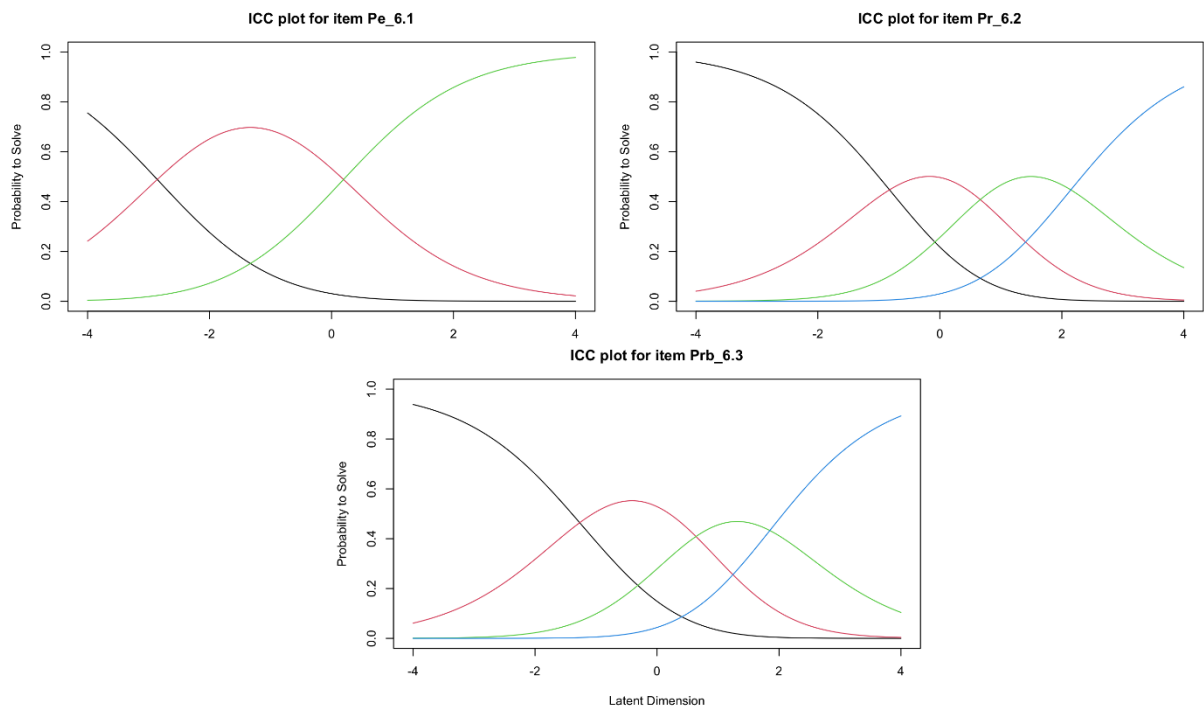


Figure 6.10 CCC plots for task 6

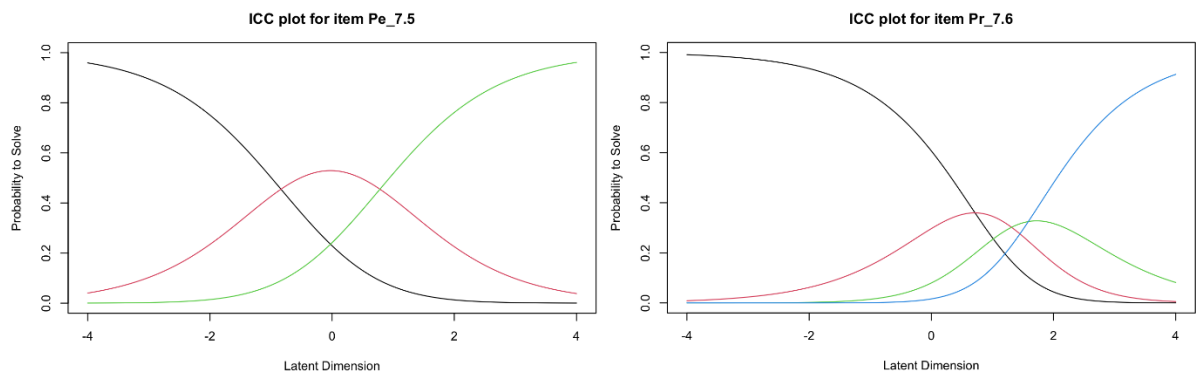


Figure 6.11 CCC plots for task 7

As for the reliability of the test, the separation reliability in the Rasch measurement theory is the ratio of the ‘true’ (observed minus error) variance to the obtained variation with values ranging between 0 and 1 (Duncan et al., 2003; Bond & Fox, 2015). The results showed that the person separation reliability was 0.77 and the item separation reliability was 0.99, which indicate reliable estimates for both person and items effects (Bond & Fox, 2015; Duncan et al., 2003).

Overall, the data fit the PCM model appropriately and most of the items have desirable characteristics. There are however still some minor problems shown in the results, such as the underfitting E-SA items and P-SA items with reversal score categories. These items will be

further discussed in section 6.4. In general, the evidence above supported that ‘*The internal structure of the construct represented in the assessment is **moderately** accurate*’.

### **6.3.2 Claim 5 There is no negative impact on the participants by implementing the assessment**

Although this assessment was conducted as an explorative piece of research and was low stakes, it is still of value to make sure that the students were not influenced negatively by taking the test. Students’ willingness in participating in the follow-up interviews and their experience of taking the test are used as evidence to help evaluate the claim.

72% and 46% of the participants were willing to participate in the follow-up interview in the fourth pilot and the main study respectively. The data collection period for the fourth pilot was at the beginning of the Autumn semester while for the main study it was at the end of the semester around the end of the year. Given high school students has limited time for activities unrelated to their school study, it is a relatively high portion of students who were willing to participate the interview. This therefore indicates that in general students were interested in the assessment and held a positive attitude.

In addition, students who participated in the follow-up interviews were encouraged to talk about their experience and feelings about taking the test. They were not prompted to give positive feedback, instead, the interviewer tried to build an equal conversation deliberately and encouraged them to express any of their negative experiences/feelings. From a validation perspective, a small amount of evidence on negative impacts may have more power to undermine the validity argument than a large amount of positive feedback may support it (Newton, 2016). Thus, instead of talking about positive feedbacks, this section focuses on some relatively negative information provided by participants to evaluate whether ‘*There is no negative impact on the participants by implementing the assessment*’. Chapter 8 will provide richer information in terms of the students’ experience of taking the test to inform the question whether ‘It is possible to have positive impacts on students’ learning by using the SAC assessment’.

Nine out of 30 students talked about the fact that they felt the test was difficult, in which 4 students expressed it in a way that seems like the assessment impacted their confidence in themselves, such as M4 who said that:

“I found there are many skills for me to improve, like my knowledge, the logical thinking, especially my language expression skill, you know, most of the time, it’s like I know in my mind, but it’s hard for me to express it and to make others understand me. I really realized that my expression ability is so poor.”

For each student who said something like this, I spent time to encourage them, give recognition to their performance, and providing them with some suggestions of how to improve SAC. There were also some other students who mentioned that the test has too many items especially the P-SA items that they had to write longer answers for, which made them felt tired after doing the test.

Overall, however it seems that the assessment didn’t bring any serious negative impact on students. This is reasonable given students knew that the assessment was low stakes. As was discussed in section 4.6, students were told about the purpose of the assessment and were not forced to take the test. However, the evidence above cannot reveal any impact on its use, rather it only focuses on students’ experience of taking the test generally. Thus, it seems appropriate to claim that *‘There was no negative impact on the participants by implementing the assessment’*.

### **6.3.3 Summary**

The evidence shown above reveals that when the assessment is evaluated from a macro-perspective, it is an imperfect but reasonable product. Two multiple-choice items showed relatively poor model-data fit, and the thresholds of two open-ended items didn’t operate as expected. But in general, most items showed satisfied characteristics, as well as the test as a whole.

Together with the validity argument for the micro-validation, the whole process of the assessment and the outcome of the assessment administered in the main study were justified to be appropriate. Therefore, the overarching claim of *‘It is possible to measure SAC from the three components and by using the SAC test results’* has been justified. But the journey of assessment development can always continue before it is perfected. Given the limited time, the process of modifying the assessment could not be implemented for another iteration, but the problems of the assessment product revealed by the evidence in the macro-validation will be discussed in the next section to further modify the product.

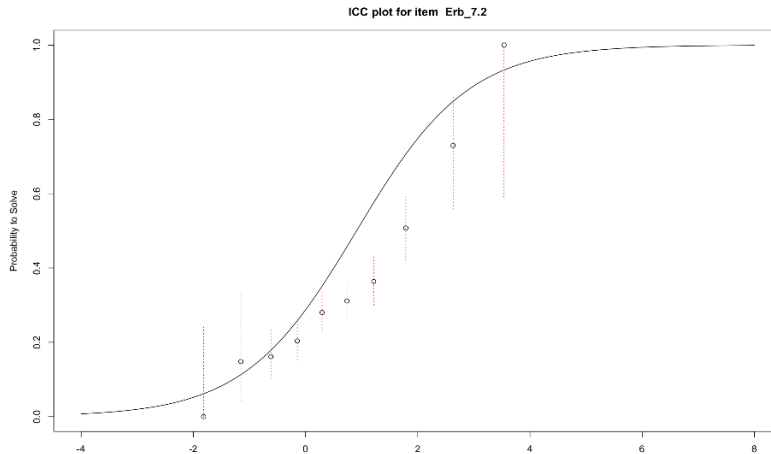
## **6.4 Further consideration about the underperforming test items**

In this section we will investigate test items that didn't perform as well as expected. These include dichotomous items that underfitted the model and polytomous items that showed underperforming category characteristics. The items will be discussed one by one according to the results shown in Table 6.5 in section 6.3.1. Decisions will be made about whether the items should be excluded from the final test version so that participants' SAC can be more accurately represented by the test outcomes.

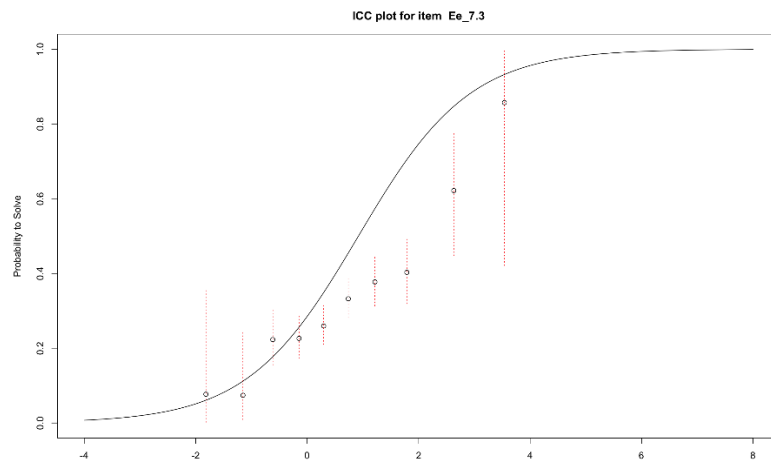
As described in section 4.5.3, empirical and expected Item Characteristic Curves (ICCs) for underfitting dichotomous items will be checked to help detect the problems of the items. For the empirical function, each dot is a raw score category and the placement of the dot on the y-axis shows how many people at that raw score actually got that item correct. If the item fits the Rasch model, the dots should roughly mirror the theoretical ICC (black curve). The red dotted lines show the 95% confidence interval for each raw score. When the bands are large, it means a less precise estimate of this raw score. In addition, when the red dotted lines do not cross the black curve, it means that it is unlikely that the empirical probabilities are matching what the Rasch model expects from the item. Category Characteristic Curves (CCCs) will be used to check the polytomous items that have reversal score categories. We will now consider the items in turn.

### **6.4.1 Items Erb\_7.2 and Ee\_7.3**

Erb\_7.2 is an Evaluation of Rebuttal item with 69% students getting it incorrect, Ee\_7.3 is an Evaluation of Evidence item with 69% students getting it incorrect. These two items show quite similar fit statistics and empirical plots, so they are analyzed together. According to the empirical plots, the overall trend of the characteristics of these items is consistent with what the Rasch model expects, but fewer students, who are located around 0 to 2 logits, get the item correct than expected by the model.



*Figure 6.12 Empirical plot for Erb\_7.2*



*Figure 6.13 Empirical plot for Ee\_7.3*

The interview data provided some potential explanation about why these items performed this way. Students who were asked about these items didn't explain clearly about how they figured these items out and few students demonstrated an adequate understanding on the item. For those who seemed ambiguous about their own thinking, they found it clear and understandable after I explained it to them. Combining this with what many students mentioned, such as "I became tired when I was doing the last task" (M11) and "task 7 provided so much information and I felt overwhelmed about analyzing it" (M2), the possible reason of the poor fit of these items is that by the time they reached this question students were fatigued, and the possibility of guessing rose as well. This seems possible given there are similar findings found in previous studies (Cheng & DeLuca, 2011). But no evidence from the interviews showed that there are specific problems in these items. However, to obtain scores that can represent the students' SAC more accurately, the items will be temporarily excluded from the final test version.

### 6.4.2 Items Prb\_4.3 and Prb\_5.3

Prb\_4.3 is a Production of Rebuttal item, there are 588 students scored 0, 478 students scored 1, 221 students scored 2, and 126 students scored 3. Category 2 for this item shows an almost invisible peak, which means this category cannot easily represent a certain level of performance and a certain level of ability of the sample. Therefore, it will be better for this category to be merged with its most close adjacent category. Category 1 is ‘students only pay attention to their own claim without analyzing other’s weaknesses, or only pointing out other’s weakness without explaining why’, and category 2 is ‘comparing his/her own argument with others, but the rebuttal as a whole is not fully reasonable (because of content knowledge) or doesn’t provide further explanation to make the rebuttal more valid’. These two categories can be combined reasonably as ‘students have the awareness of providing rebuttal and have some understanding about rebuttal, but their rebuttal is not rigorous or sufficient enough to fully weaken other’s argument and justify their own’. After merging these two categories, the CCC plot becomes more acceptable, and the threshold distance increases from 0.9 logits to 2.4 logits that is more desirable.

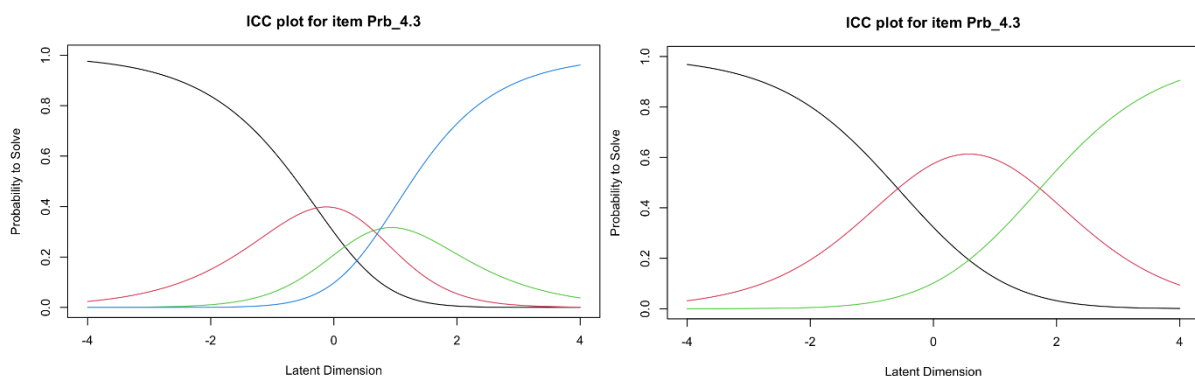


Figure 6.14 CCC for Prb\_4.3 (before and after revision)

In a similar way to item Prb\_4.3, Prb\_5.3 is a Production of Rebuttal item, and this time category 1 doesn’t appear to be the most likely observation anywhere along the latent trait continuum. There are 596 students scored 0, 371 students scored 1, 427 students scored 2, and 19 students scored 3. In addition, there is reversal of thresholds 1 and 2, which means that getting a score of 2 is easier than getting a score of 1. By checking the scoring rubrics and students’ test papers, it was found that a score of 2 was usually given to answers that merely pointing out the provided counterevidence without further explaining it. So, it should be reasonable to merge category 2 with category 1 as partial valid rebuttal.

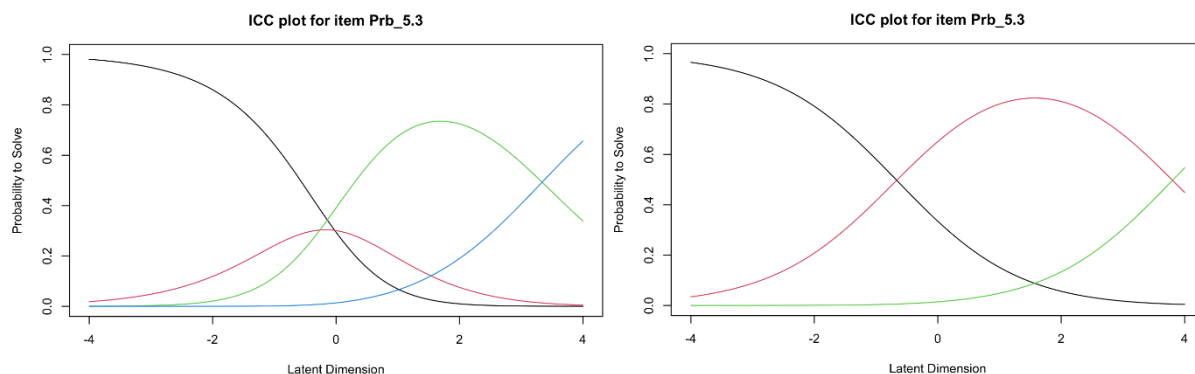


Figure 6.15 CCC for Prb\_5.3 (before and after revision)

For both items, students in the interview mentioned that they were not sure about how to rebut other's argument. However, it turned out that when obvious counterevidence is provided in a task, it is much easier for students to recognize and use it to form their rebuttal. Nevertheless, for the Prb item in Task 6 (SSI task), some students mentioned that 'there is no correct or wrong for this topic...as long as what I said is reasonable. So, I just need to support my own claim' (M7). It seems that the students' attending to other's argument was highly dependent on the context, particularly whether they viewed it necessary to attend to other's argument and how easy it was to do so.

### 6.4.3 Summary

In summary, after further checked the underperforming items, the final Rasch score will be calculated by excluding the scores of Erb\_7.2 and Ee\_7.3 and by modifying the score categories of Prb\_4.3 and Prb\_5.3. Combining with the validity arguments demonstrated in section 6.2 and section 6.3, it is justifiable to say that the Rasch score resulted from further modifying these items is an accurate measure of students' SAC. So, it is reasonable for this study to use the final Rasch score for statistical analysis in Chapter 7.

In addition, although Erb\_7.2 and Ee\_7.3 were excluded from the final interpretation of students' SAC, this is mainly because of the consideration that most students felt tired on the last task and the probability of guessing increased, rather than saying they are actually bad items since no evidence was obtained that indicate they have problems. Moreover, the test still includes all the SAC elements after excluding these two items.

### 6.5 Further consideration about the construct of SAC

By further modifying the items in section 6.4, a final set of data (i.e., Rasch scores after



modifying underperforming items) that can accurately represent the students' SAC was obtained for further statistical analysis. This section will further consider the assessment results (shown in Table 6.6) after modifying the score categories of Prb\_4.3 and Prb\_5.3 to expand upon our understanding of the SAC construct. The assessment results used in this section do not exclude Erb\_7.2 and Ee\_7.3 as we will be doing in Chapter 7 because

- 1) they each represents important SAC elements,
- 2) their MSNQ fit statistics do not deviate much from the normal range,
- 3) no problems of the items were revealed from the interview.

### **6.5.1 Implications for a learning progression of SAC**

Table 6.2 in Claim 1-3 indicated that the overall pattern of the items estimated in the fourth pilot was consistent with the general assumption (i.e., I-SA is the easiest, followed by E-SA and P-SA, and it is relatively easier to generate simple argument), nevertheless,

- 1) some P-SA items were easier than expected, while E-SA and even I-SA items were harder than assumed; and
- 2) the difficulty of the SAC elements in each component was not consistent with the assumption (i.e., Evidence was the easiest, followed by Reason and Rebuttal).

However, after figuring out the possible problems in items and modifying them, this inconsistency still exists in the final test version, as shown in Table 6.6. So, it is necessary to reflect on the theoretical assumptions of the SAC construct. The shared patterns in the results are that:

- 1) the complexity of E-SA items is not exactly below P-SA items, but between the highest (a score of 2) and lowest (a score of 0) proficiency level of P-SA items;
- 2) the three SAC components do not seem to be precisely in a linear progressive relationship, in other words, the SAC elements in each component don't represent the same level of SAC.

Table 6.6 Items estimates (after modifying Prb\_4.3 and Prb\_5.3)

Item	Location(logit) (SE)	Threshold 1(logit) (SE)	Threshold 2(logit) (SE)	Threshold 3(logit) (SE)
Ee_2.1	-2.97 (0.10)			
I_1	-1.81 (0.07)			
Erb_3.4	-1.41 (0.06)			
Pe_6.1	-1.32	-2.86 (0.11)	0.21 (0.06)	
Pe_4.1	-0.74	-2.11 (0.08)	0.63 (0.06)	
Er_7.4	-0.30 (0.06)			
Ee_3.1	-0.14 (0.06)			
Erb_2.2	-0.05 (0.06)			
Pe_7.5	0.01	-0.83 (0.06)	0.81 (0.07)	
Ie_3.2	0.02 (0.06)			
Ir_7.1	0.09 (0.06)			
Pr_4.2	0.22	-0.24 (0.07)	0.27 (0.08)	0.62 (0.10)
Pe_5.1	0.27	-1.47 (0.06)	2.00 (0.10)	
Prb_6.3	0.43	-1.26 (0.07)	0.66 (0.07)	1.88 (0.14)
Erb_7.2	0.56 (0.06)			
Ee_7.3	0.57 (0.06)			
Prb_4.3	0.58	-0.58 (0.06)	1.73 (0.10)	
Pr_6.2	0.68	-0.82 (0.06)	0.68 (0.07)	2.18 (0.17)
Er_3.3	0.82 (0.06)			
Pr_5.2	1.14	0.40 (0.06)	0.74 (0.09)	2.28 (0.20)
Pr_7.6	1.23	0.73 (0.06)	1.34 (0.11)	1.63 (0.19)
Pr_5.3	1.57	-0.67 (0.07)	3.80(0.08)	

Although it was assumed that it may be easy for high school students to generate simple argument, as mentioned in section 5.2.1, empirical evidence is needed to check the detailed difficulty order of these items to expand our understanding of SA. For the I-SA items, as described in section 5.4, Ic and Irb items showed extremely low difficulty in the pilots, and Table 6.6 shows that Ie and Ir items have higher estimates than expected. Similarly, previous studies argued the key role of differentiating between claim, evidence, and reason, and found that students have difficulties in differentiating between them and usually fail to articulate them in an argument (Berland & Reiser, 2009; Driver et al., 2000; Kuhn & Reiser, 2005). Thus, it seems reasonable to assume that identifying *evidence* and *reason* are more demanding than identifying *claim* and *rebuttal* in an argument.

When checking the E-SA items, it seems that the content knowledge embodied in the items and the nature of the SA element assessed by the items affected the items' difficulty. Specifically, Ee items (i.e., Ee\_3.1 and Ee\_7.3) in which more content knowledge is needed to make evaluation have higher difficulty estimates compared with items (i.e., Ee\_2.1) in which *evidence* is superficially connected with the *claim*. This is consistent with previous studies in which students with high content knowledge proficiency tend to engage better in SA (Yang et

al., 2015; Liu et al., 2019).

As for Erb items, it seems difficult for students to evaluate when the *rebuttal* is explicitly analysing or critiquing the opposing argument (i.e., Erb\_2.2 and Erb\_7.2) rather than expressing their disagreement by simply putting forward their own argument (Erb\_3.4). Previous studies also revealed that students have difficulties and are unaware of attending to other's arguments (Deng & Wang, 2017; Chen et al., 2019). Er items turn out to be generally more difficult for students. This seems reasonable given its nature is to connect between *claim* and *evidence*, more content knowledge (i.e., making the connection accurate) and the ability to explain (i.e., making the connection coherent) is needed to build this connection (Berland & Reiser, 2009). Thus, based on the items design of this study, it seems that the difficulty of the Ee items is decided by the complexity of the connection between *evidence* and *claim*; Er items are generally more demanding without more nuanced findings; the difficulty of Erb items is decided by whether the *rebuttal* attends to others' argument.

In terms of the P-SA items, the results show that generating relevant and simple SA is easier for high school students especially when less content knowledge is needed (i.e., social scientific task-Task 6). For scientific tasks, generating SA is easier when the context is less complex (i.e., less content knowledge, or less information to compare, or familiar topic) with the general difficulty of Task 6 < Task 4 < Task 5/Task 7. Moreover, Pe items are easier compared to other P-SA items. The above results are generally consistent with previous studies that students can construct argument from an early age, providing evidence is relatively easier and content knowledge and provided information influence the difficulty of engaging in SA (Kuhn et al., 2010; McNeill et al., 2006; Deng et al., 2017; Berland & Reiser, 2009; Berland & McNeill, 2010), as discussed in Chapter 3. Overall, choosing relevant *evidence* and generating simple *reason* and *rebuttal* seem less demanding than choosing sufficient *evidence* and generating plausible *reason* and *rebuttal*, and constructing *reason/rebuttal* that is accurate in the content knowledge and coherent in its logic still seems challenging for those students.

From a broader view of looking at all the three SAC components, firstly, the ability needed to judge a given argument according to the provided criteria is higher than to construct a simple argument, but lower than to generate a well-constructed argument, which is consistent with what participants said in the interview that “the evaluation items are difficult...but the open-ended items, you know, it is easy to at least say something” (M5). Specifically, there were 6 students who mentioned that they found both the E-SA items and P-SA items difficult although

in different ways. Four of these 6 students mentioned that “it was hard to form a good argument to support my claim” (M3), and they were not confident about the argument they constructed although “it is not difficult to at least say something relevant (to my claim)” (M7).

Secondly, it is generally easier for the students to provide or choose between relevant evidence than to provide effective evaluation. Thirdly, it is not as easy as expected for the students to identify SAC elements from a context of argumentation, especially to identify *evidence* and *reason*. Taking the above considerations together, A potential learning progression of SAC is therefore generated as shown in Table 6.7.

Table 6.7 Learning progression of SAC


Components	Simple  Complex		
	Level 1	Level 2	Level 3
<b>I-SA</b>	I-1 Students can identify claim or rebuttal in an argument.	I-2 Students can identify reason or evidence in an argument.	
<b>E-SA</b>	E-1 Using the provided criteria, students can evaluate evidence and rebuttals that are less cognitively demanding: 1) recognize whether evidence is relevant or sufficient to support the claim when the evidence and the claim is connected/unconnected in a direct and superficial way; 2) recognize that the rebuttal does not explicitly attend to other’s argument.	E-2 Using the provided criteria, students can evaluate evidence and rebuttals that are more cognitively demanding: 1) when there are implicit logical chains (content knowledge is needed) between the evidence and the claim; 2) when rebuttal engages in other’s argument. 3) Students can evaluate reason.	
<b>P-SA</b>	P-1 1) Students can identify relevant evidence from the provided information pool; and 2) can generate relevant and simple rebuttal and reason in social scientific issue (SSI) that does not need content knowledge.	P-2 1) Students can identify all the relevant evidence from the provided information pool; and 2) can generate comprehensive reason and rebuttal (but lack of scientific accuracy or lack of coherence) in scientific context.	P-3 Students can generate coherent and accurate reason and rebuttal under SSI and scientific context.

Figure 6.16 below visually presents the reconsideration of the SAC construct based on the assessment results of this study (As the colour of the band deepens, the item/threshold it represents becomes more difficult. The lightest bands represent items/thresholds at level 1, the darkest bands represent items/thresholds at level 3). Items/thresholds with higher difficulty and students demonstrating higher level of SAC are located toward the right side of the continuum.

It seems that the reconsideration of the SAC construct indeed indicates a possible learning progression of SAC. The next section will explore how the group of Chinese high school students perform on SAC based on the proposed learning progression.

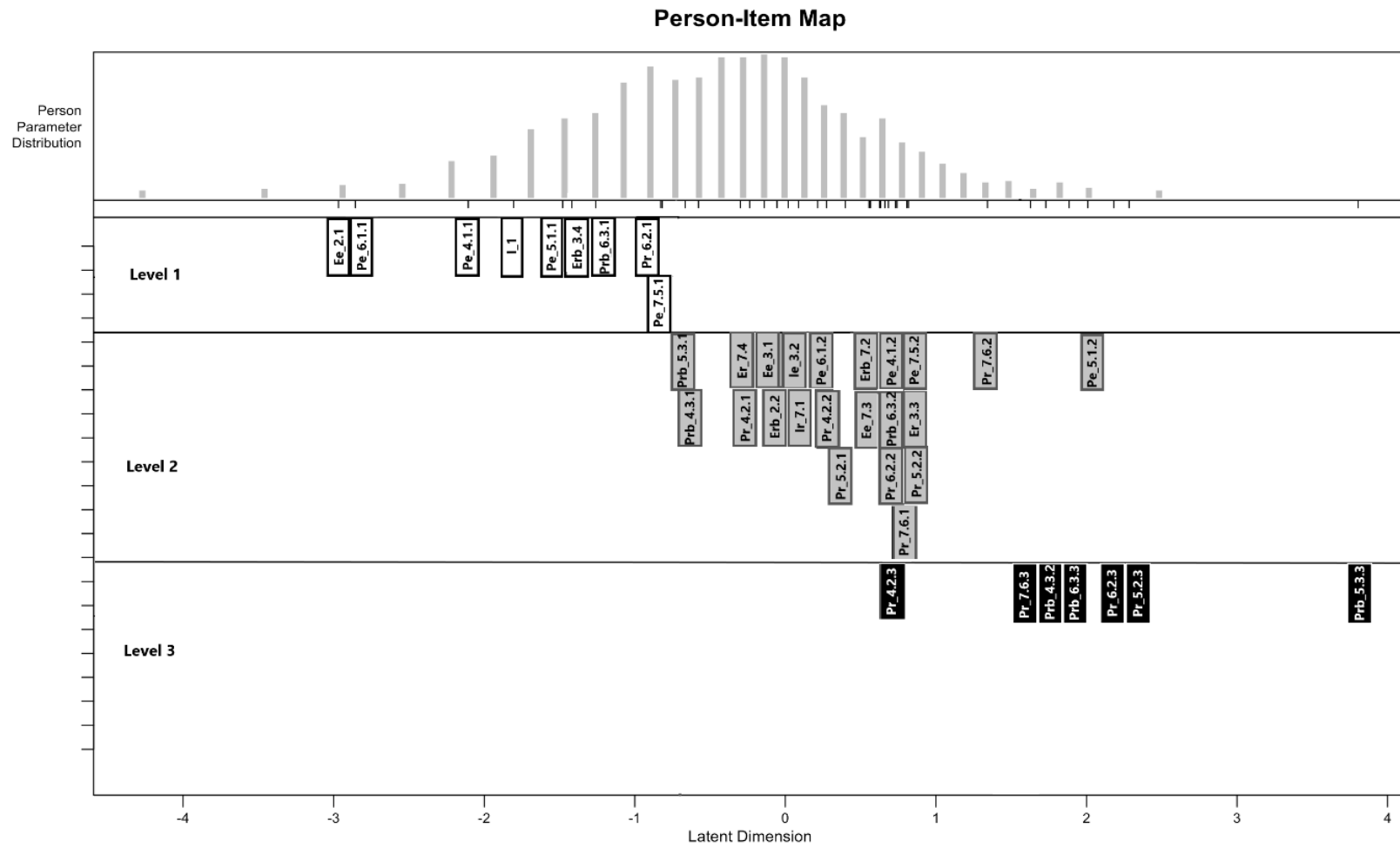


Figure 6.16 Wright map showing the progression levels of SAC

## 6.5.2 Chinese high school students' performance on the SAC learning progression

Based on the expanded understanding of SAC as a learning progression, it seems that the Rasch score generated can be given more of an interpretation in terms of competence levels. To assign students into corresponding progression levels based on their Rasch scores, the average of the item/threshold estimates within each level is used as the cutoff point between SAC levels as shown in Table 6.8 (Shi et al., 2021). In detail, the average difficulty estimates of items or item thresholds (e.g., Pe\_6.1.1 indicates obtaining a score of 1 on item Pe\_6.1) are calculated and those that are within the same level are presented in Table 6.8. Students who have the ability estimates that are lower than the average value would be taken as not reaching the required competence entailed by that level.

*Table 6.8 Measures for each SAC level*

Level	Item/Item threshold	Average measure (logit)	Measure range (logit)
Level 0			$(-\infty, -1.73)$
Level 1	I_1, Ee_2.1, Erb_3.4, Pe_6.1.1, Pe_4.1.1, Pe_7.5.1, Pe_5.1.1, Prb_6.3.1, Pr_6.2.1	-1.73	$[-1.73, 0.39)$
Level 2	Pe_6.1.2, Pe_4.1.2, Er_7.4, Ee_3.1, Erb_2.2, Pe_7.5.2, Ie_3.2, Ir_7.1, Pr_4.2.1, Pr_4.2.2, Prb_6.3.2, Erb_7.2, Ee_7.3, Pr_6.2.2, Er_3.3, Pe_5.1.2, Pr_5.2.1, Pr_5.2.2, Pr_7.6.1, Pr_7.6.2, Prb_4.3.1, Prb_5.3.1	0.39	$[0.39, 2.02)$
Level 3	Pr_4.2.3, Prb_6.3.3, Prb_4.3.2, Prb_5.3.2, Pr_6.2.3, Pr_5.2.3, Pr_7.6.3	2.02	$[2.02, +\infty)$

To illustrate the characteristics of each level, a Rasch measure below -1.73 indicates level 0 of SAC, where students has not reached the average measure of items in level 1. Students at level 0 demonstrate nearly no (even basic) understanding of SA and they do not possess the skills contained in Level 1 of SAC. In more detail, students at this level cannot recognize the claim and rebuttal in an argument, and it is also difficult for them to generate simple arguments. In addition, they tend to lack the basic epistemic understanding about argumentation in science.

A Rasch measure between -1.73 to 0.39 represents level 1 of SAC, indicating a person who has reached or outperformed the average measure of items in level 1 but has not reached that of the average of the level 2 items. At this level, students have a basic understanding about SA and can recognize the claim and rebuttal in an argument. When provided with the criteria, students can evaluate evidence in situations where the connection between evidence and claim is direct. Students can also recognize rebuttals that are not attending to the opposing argument. To put

it simply, students have a basic epistemic understanding of SA, and can apply this understanding to evaluate arguments when the context is not complex. In terms of generating their own argument, students at this level are trying to produce statements that are relevant to the problem to be argued, they demonstrate that they know they are constructing an argument. But their arguments are not targeting the problem being argued in the right direction and are inaccurate and incoherent.

If a student gets a Rasch score between 0.39 to 2.02, the SAC of this student is at level 2. At this level, students can deal with argumentation tasks that are more complex, and they possess a deeper understanding about SA. Students understand and can identify evidence and reason in an argument; can evaluate evidence that connects with a claim in a more complex way and can evaluate a rebuttal in terms of whether it weakens the opposing argument. Furthermore, students at this level can produce a plausible argument, although with some inaccurate content knowledge understanding or incoherent expression. Generally, students at this level know what they are doing and have a sense of how to do it: they have more epistemic understanding about SA, and they demonstrate the ability to apply their understanding in evaluating and generating scientific arguments in the right direction.

If a student gets a Rasch score above 2.02, this student can produce accurate and coherent arguments, and has the potential to engage in more complex argumentation. Students at level 3 know what they are doing and how to do it, and they can do it well.

To understand how the students in this study perform on each SAC level, students' performance on the test are categorized into different SAC levels based on their Rasch scores (the distribution of their SAC Rasch scores is shown in Figure 6.17). Most of students (72%) who participated in the study scored in a range that is categorized as being at level 1, while only 19.9% of the participants were at level 2. Although a small portion of participants were at level 0 (7.8%), only 0.3% of the sample were at level 3. The results indicate that most of the students in the sample have some basic understanding of SA and can generate simple arguments, and some of the students can generate more accurate and coherent SA and have certain epistemic understanding of SA but they still need more instructions to become more proficient at engaging in SA.



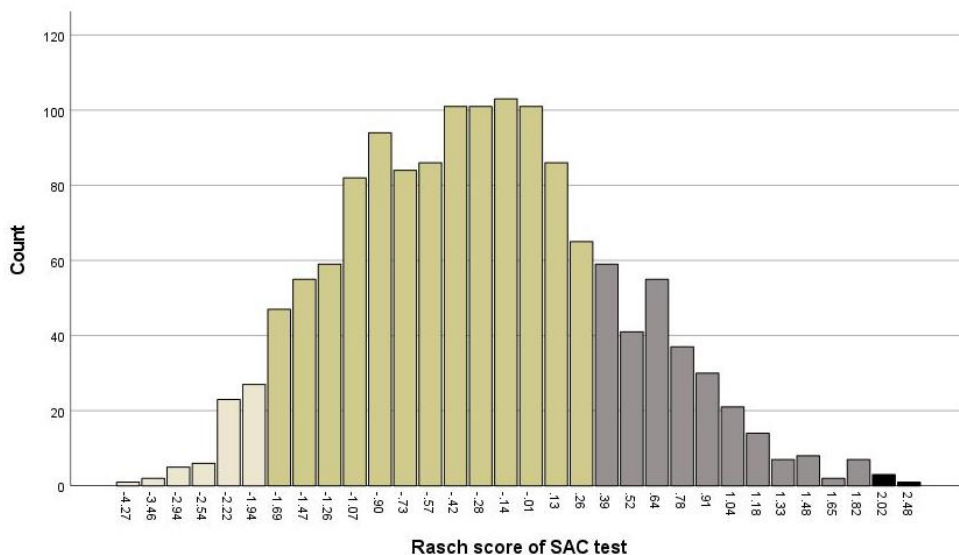


Figure 6.17 Distribution of the students' SAC

## Chapter summary

This chapter answered RQ 2 and RQ 3. For RQ 2, the SAC assessment was validated from both a micro and macro perspective. An interpretation argument for the assessment (IUA), including five claims, was proposed in section 6.1 to guide the validation. Each claim in the IUA was evaluated in section 6.2 and section 6.3. As a result, the micro-claim was justified as *'The assessment procedure was conducted in a **moderately** appropriate and effective way and elicited participants' SA'*, with a potential weakness of the micro validity argument that the administration process was not well controlled due to failing to monitor the administration in person due largely to the pandemic. The macro-validation appropriately supported the overarching assessment claim that *'It is possible to measure SAC from the three components and by using the SAC test results'* however some items showed poor performance. Overall, the overarching assessment claim has been **appropriately** (but not perfectly) supported from both perspectives, thus the assessment results can represent the students' SAC performance.

The underperforming items revealed by the evidence in the macro-validation were further checked by referring to the empirical and expected ICC and the participants' interview data in section 6.4. Consequently, score categories of Prb\_4.3 and Prb\_5.3 were modified to show satisfactory performance and Erb\_7.2 and Ee\_7.3 were excluded when calculating the final Rasch score that can be used in further statistical analysis.

For RQ 3, the assessment results after modifying Prb\_4.3 and Prb\_5.3 were reconsidered in section 6.5 to inform an expanded understanding of SA. By analysing the items and

corresponding item estimates, a three-level learning progression of SAC was proposed. The learning progression revealed how the characteristics of SA element and assessment task influenced the demands of evaluating and constructing an argument. This understanding of SAC as a learning progression brought an expanded interpretation for the assessment results, namely, the resulted Rasch scores can be categorized into three different levels of SAC. It turned out that most of the participants were currently at lower levels of SAC.

Overall, this chapter offered more understanding about the construct of SAC and the participants' performance on SAC by validating the SAC assessment and analysing the SAC assessment results. The next chapter will further promote the understanding of SAC and Chinese high school students' SAC performance by exploring the relationships between the students' SAC performance and relevant factors.

## **Chapter 7. Students' SAC Performance and Relevant Factors**

### **Introduction**

This chapter aims to answer Research Question 4 to understand what influenced the students' performance on the SAC assessment. The SAC Rasch scores, obtained by applying a PCM model, on the SAC test after modifying scores for items Prb\_4.3 and Prb\_5.3 and excluding items Erb\_7.2 and Ee\_7.3 are used to represent the students' performance on the SAC assessment (see section 6.4). To be specific, section 7.1 will briefly introduce the data subsets that are used in this chapter. Section 7.2 will investigate whether the students in different classes, schools, areas and of different genders showed different performance on the assessment. This informs our understanding about whether the students' learning environment may affect their SAC and thus how generalizable our findings are. Section 7.3 will explore whether providing the students with the definition of SA promoted their performance on the test given SA was unfamiliar to the students. Section 7.4 looks into links between their content knowledge proficiency and SAC.

### **7.1 Data subsets**

This section introduces the data subsets collected in the main study that are involved in the statistical analysis in this chapter. The information in terms of the classes, schools, and areas where the participants are located has been described in section 4.3.4. As mentioned in section 4.4.2.5, only a subset of data was involved when considering scaffolding and test papers with scaffold (i.e., definition of SA) were randomly assigned to the students within classes and the classes were also randomly selected to assign these two versions of test papers. In total, there were 618 students in 15 classes that were assigned with either version of test paper, in which 332 students did the test with scaffolding and 286 students did the test without scaffolding. The unbalanced sample size for the two groups was due to the fact the students had the right to not respond to or submit the test paper (see section 4.6).

As also mentioned in section 4.4.2.5, only three schools provided the school Physics test scores of their students, of which one school provided both Physics and Chinese test scores. School 3 provided the Physics and Chinese school test scores of 393 students, school 5 provided the Physics school test scores of 126 students, and the school Physics test scores of 190 students in school 6 were collected. In terms of gender, there were in total 701 (50%) boys and 665 (47%) girls participating in the main study with 47 students who didn't report their gender.

## 7.2 Context and gender

### 7.2.1 Area

518 of the participants studied in Shenzhen city (area 1) and 895 of the participants studied in Jilin province (area 2). Multilevel models are used to analyze the difference in SAC performance of students in the two areas whilst accounting for the fact that students are nested within classes. As shown in Table 7.1, Model 1 is a two-level variance-components model for SAC Rasch scores without covariates, while Model 2 adds the area that the students were located in as a covariate in the model. The Variance partition coefficient (VPC) is used to represent how much of the variance in students' performance can be explained by the clustering due to classes (Goldstein, Browne, & Rasbash, 2002). Model 1 shows that the  $VPC = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = 0.30$ , indicating that 30% of SAC student score variation lies between classes, and 70% within classes. There is substantial clustering or non-independence in the data thus it is appropriate to use the multilevel model.

Moving from Model 1 to Model 2 results in a better model with LR (likelihood ratio test statistic) = 6 ( $p = .014$ ), and the variation explained by classes reduces to 27% as part of it (3% of total variation) is now explained by classes being nested in particular areas. Model 2 reveals that students in Jilin scored 0.426 lower than those in Shenzhen, and the difference is significant given  $p = 0.009$ . So, it seems that on average students in Shenzhen city performed better on the SAC assessment than students in Jilin province. Although the result seems reasonable given Shenzhen is a big city while Jilin is a remote province in the northeast of China, no conclusions can be drawn in terms of the two regions since the students were based on only several schools and were not randomly recruited across the regions. So, examining the performance differences between schools may help better understand the difference in the two regions.

Table 7.1 Area difference in SAC performance

Parameter	Model 1		Model 2		Model 3	
	Estimate	SE	Estimate	SE	Estimate	SE
<b>Fixed-part</b>						
Intercept	-0.378	0.087*	-0.129	0.124	0.051	0.117
Area 2			-0.426	0.162*		
School 2					-0.585	0.213*
School 3					-0.506	0.165*
School 4					-0.613	0.291*
School 5					-0.507	0.250*
School 6					-0.436	0.223
School 7					-1.651	0.303*
<b>Random-part</b>						
Student variance ( $\sigma_e^2$ )	0.618	0.024	0.618	0.024	0.618	0.024
Class variance ( $\sigma_u^2$ )	0.268	0.065	0.224	0.056	0.132	0.034
<b>Deviance</b>	3437		3431		3413	
<b>VPC</b>	0.30		0.27		0.18	
<b>Number of classes</b>	38		38		38	
<b>Number of students</b>	1413		1413		1413	

Notes: \* $p < .05$ .

### 7.2.2 School

In a similar way to has been illustrated in the previous section, the school attended by the students is added to the multilevel model as a covariate in place of area. Moving from Model 1 (a two-level variance-components model for SAC Rasch score without covariates) to Model 3 (add school attended as a covariate in the model) results in a better model with LR = 24 ( $p < .001$ ), and the variation explained by classes reduces from 30% to 18% as much of it is due to classes being nested in particular schools. So, school effects contribute a notable part (12% of total variance) in explaining the variance in students' SAC performance.

The comparison between pairs of schools is illustrated in Table 7.2 below. The results show that, except for school 6, the average SAC performance of the students in school 1 is significantly higher than that of other schools. The average SAC performance of the students in school 7 is significantly lower than that of other schools. However, the average SAC performance of the students in schools 2, 3, 4, 5, and 6 are not significantly different from each other. School 1 is in Shenzhen city while school 7 is in Jilin province. So, the area difference found in the previous section maybe due to the outstanding performance of the students in school 1. The results are not surprising, since school 1 is a high-achieving high school whose proportion of students entering good universities in recent years is slightly lower than that of the traditional 10 elite high schools in Shenzhen. School 5 has the highest achievement among

the invited schools in Jilin, but its performance is lower than school 1.

*Table 7.2 School difference in SAC performance*

	<b>School 1</b> (N=385)	<b>School 2</b> (N=133)	<b>School 3</b> (N=401)	<b>School 4</b> (N=119)	<b>School 5</b> (N=128)	<b>School 6</b> (N=195)
<b>S 1</b>						
<b>S 2</b>	- 0.585(0.213*) )					
<b>S 3</b>	- 0.506(0.165*) )	0.080(0.212)				
<b>S 4</b>	- 0.613(0.291*) )	-0.028(0.320)	-0.107(0.291)			
<b>S 5</b>	- 0.507(0.250*) )	0.078(0.283)	-0.002(0.250)	0.106(0.347)		
<b>S 6</b>	-0.436(0.223)	0.149(0.260)	0.069(0.223)	0.177(0.328)	0.071(0.292)	
<b>S 7</b>	- 1.651(0.303*) )	- 1.066(0.331*) )	- 1.146(0.303*) )	- 1.039(0.386*) )	- 1.144(0.356*) )	- 1.215(0.338*) )

Notes: \* $p < .05$ .

### 7.2.3 Class

As the previous section showed that the effect of classes explained a substantial part of the SAC score variances in the group of participants. So, in this section we will have a closer look at whether it is the class type (i.e., key classes in which students have high performance or ordinary classes, see section 4.3.4) that has influenced the students' SAC performance. This section uses multilevel models to test whether class type could explain the variance in the students' performance by considering all the schools at first, then to test how the class type influenced the schools differently.

There were in total 455 students from the key classes in these 7 schools. As shown in Table 7.3, Model 4 adds class type as a covariate and Model 5 adds the interactions between key classes and schools in the model. It has been shown previously that school effects also play a role in explaining the variance in students' SAC performance. Moving from Model 3 to Model 4 leads to a better model with  $LR = 13$  ( $p < .001$ ), Model 4 reveals that the students in key classes on average scored 0.474 higher than those in ordinary classes, and the difference was significant given  $p < .001$ . So, in general, class type influenced the students' performance on

the SAC assessment.

However, moving from Model 4 to Model 5 does not quite result in a better model given LR = 8 ( $p = .16$ ). So, the difference in SAC performance between the two class types does not vary significantly between schools.

Table 7.3 Class type and SAC performance

Parameter	Model 3		Model 4		Model 5	
	Estimate	SE	Estimate	SE	Estimate	SE
<b>Fixed-part</b>						
Intercept	0.051	0.117	-0.084	0.104	-0.054	0.104
School 2	-0.585	0.213*	-0.648	0.180*	-0.833	0.205*
School 3	-0.506	0.165*	-0.583	0.140*	-0.675	0.157*
School 4	-0.613	0.291*	-0.714	0.243*	-0.786	0.293*
School 5	-0.507	0.250*	-0.527	0.209*	-0.274	0.224
School 6	-0.436	0.223	-0.302	0.189	-0.332	0.174
School 7	-1.651	0.303*	-1.516	0.258*	-1.546	0.236*
Key class			0.474	0.120*	0.362	0.194
Key class*school 2					0.477	0.335
Key class*school 3					0.254	0.261
Key class*school 4					0.196	0.436
Key class*school 5					-0.755	0.400
<b>Random-part</b>						
Student variance ( $\hat{\sigma}_e^2$ )	0.618	0.024	0.619	0.024	0.619	0.024
Class variance ( $\hat{\sigma}_u^2$ )	0.132	0.034	0.087	0.024	0.066	0.019
<b>Deviance</b>	3413		3400		3392	
<b>VPC</b>	0.18		0.12		0.10	
<b>Number of classes</b>	38		38		38	
<b>Number of students</b>	1413		1413		1413	

Notes: \* $p < .05$ .

### 7.2.4 Gender

This section explores whether boys and girls performed differently on the test. As shown in Table 7.4, Model 1 is a two-level variance-components model for SAC Rasch score without covariates, while Models 3 and 6 add school attended and gender as covariates in the model respectively. After accounting for both schools and classes, Model 6 reveals that girls scored 0.021 lower than boys, but the difference was not significant given  $p = 0.626$ . But who didn't report their gender scored 0.335 lower than boys, which is significant given  $p < .01$ . So, it is concluded that there is no significant gender difference in the students' SAC performance. The significantly lower performance of students who didn't report gender maybe due to that they didn't take the assessment seriously since they even didn't provide complete background information.

Table 7.4 Gender difference in SAC performance

Parameter	Model 1		Model 3		Model 6	
	Estimate	SE	Estimate	SE	Estimate	SE
<b>Fixed-part</b>						
Intercept	-0.378	0.087*	0.051	0.117	0.072	0.121
School 2			-0.585	0.213*	-0.583	0.217*
School 3			-0.506	0.165*	-0.512	0.169*
School 4			-0.613	0.291*	-0.601	0.297*
School 5			-0.507	0.250*	-0.504	0.255
School 6			-0.436	0.223	-0.446	0.227
School 7			-1.651	0.303*	-1.667	0.308*
Girl					-0.021	0.043
No gender report					-0.335	0.124*
<b>Random-part</b>						
Student variance ( $\sigma_e^2$ )	0.618	0.024	0.618	0.024	0.615	0.023
Class variance ( $\sigma_u^2$ )	0.268	0.065	0.132	0.034	0.132	0.034
<b>Deviance</b>	3437		3413		3405	
<b>VPC</b>	0.30		0.18		0.18	
<b>Number of classes</b>	38		38		38	
<b>Number of students</b>	1413		1413		1413	

Notes: \* $p < .05$ .

In general, students in higher-achieving schools or classes had better SAC performance than those in lower-achieving schools or classes. The result is within expectation given students with high school achievement may have higher motivation to engage in the test and with higher content knowledge proficiency and cognitive ability. Although previous studies tended to report no gender difference in students' SA performance (Deng, 2015; Kuhn, 1991), gender difference in learning especially in science area has been discussed a lot among Chinese society and scholars. The opinion that boys have higher talent in learning science has been prevailing in Chinese society, an impression shaped and reinforced by cultural and social expectations (Wu & Guo, 2019). PISA 2015 and PISA 2018 reported that Chinese girls had lower performance than boys in math and science (OECD, 2020; Wu & Guo, 2019). Although girls have outperformed boys in recent years in Gaokao and girls perform similarly with boys among high science performance students (Shao & Pang, 2016; Hu & Tang, 2013), girls were found to have lower interest, confidence, and motivation on learning science (Tang & Hu, 2013; Kuang, 2019; Wu & Guo, 2019). However, this study didn't find significant gender difference in SA engagement among the students participated the test. As discussed in Chapter 2 that students in their second year of high school have to choose to study science/Physics or not, the result thus may be due to that the sample in this study only includes girls that choose to study Physics, who may have already been more interested or motivated to learning science.



### 7.3 Scaffold

To further understand students' SAC, this section explores how much influence providing the students with the definition of SA elements (i.e., scaffolding) has on their SAC performance, especially considering students were unfamiliar with SA and they talked in interviews about how scaffolding helped them (see section 8.3). As shown in Table 7.5, Model 3 accounts for schools, Model 7 adds whether a student took the test paper with scaffolding as a covariate, and Model 8 further adds whether a class were using the test paper with scaffolding (2), without scaffolding (0), or both (1) as covariates in the model. There should have been no classes that provide all students with test papers without scaffold, and this was caused by teachers' inappropriate operations (see section 4.4.2.5).

Model 7 turns out to be similar to Model 3 given  $LR = 0$ . Students who took the test paper with scaffolding scored 0.014 lower than those who took the test paper without scaffolding, and this result is not significant. Moving from Model 3/7 to Model 8 does not result in a better model given  $LR = 4$  ( $p = .26$ ). After accounting for both schools and classes, Model 8 reveals that the classes that were administered the test paper with scaffolding scored 0.365 higher than those administered the test paper without scaffolding, while the classes that were administered both kinds of test paper scored 0.530 higher than those administered the test paper without scaffolding. Students who took the test paper with scaffolding scored 0.019 lower than those who took the test paper without scaffolding. But all these differences are not significant ( $p = 0.058$ ,  $p = 0.159$ ,  $p = 0.746$  respectively). So, whether a class was provided with the test paper with scaffolding had no influence on the students' SAC performance, and interestingly the scaffold provided in the test had no influence on the SAC performance of the students.

Table 7.5 Scaffold and SAC performance

Parameter	Model 3		Model 7		Model 8	
	Estimate	SE	Estimate	SE	Estimate	SE
<b>Fixed-part</b>						
Intercept	0.051	0.117	0.061	0.125	-0.329	0.257
School 2	-0.585	0.213*	-0.585	0.213*	-0.658	0.207*
School 3	-0.506	0.165*	-0.504	0.166*	-0.535	0.159*
School 4	-0.613	0.291*	-0.616	0.292*	-0.754	0.291*
School 5	-0.507	0.250*	-0.508	0.250*	-0.589	0.242*
School 6	-0.436	0.223	-0.437	0.224	-0.317	0.222
School 7	-1.651	0.303*	-1.655	0.304*	-1.450	0.309*
Classwithscaffold_1					0.530	0.279
Classwithscaffold_2					0.365	0.259
Scaffold			-0.014	0.058	-0.019	0.060
<b>Random-part</b>						
Student variance ( $\sigma_e^2$ )	0.618	0.024	0.618	0.024	0.618	0.024
Class variance ( $\sigma_u^2$ )	0.132	0.034	0.132	0.034	0.118	0.031
<b>Deviance</b>	3413		3413		3409	
<b>VPC</b>	0.18		0.18		0.16	
<b>Number of classes</b>	38		38		38	
<b>Number of students</b>	1413		1413		1413	

Notes: \* $p < .05$ .

## 7.4 Content knowledge

Content knowledge is measured in this study by using the students' scores on their school achievement tests. The relationships between the students' SAC score and school achievement test score are analyzed separately for each school because schools used different tests. For each school, the most recent formal school examination test result when the students took the SAC test was collected to represent their content knowledge proficiency. However, the specific content knowledge included in the SAC test is about 'motion and forces' while the school tests in those schools were mostly about Electromagnetism since the students were learning about that module at that time. So, the school Physics test results may represent more the students' content knowledge of Electromagnetism, or at most the students' overall Physics learning proficiency at that time but may not represent the students' proficiency on 'motion and forces' specifically.

### 7.4.1 School 5

A scatterplot is used to obtain visual information in terms of how students' SAC scores relate to their school achievement test scores. Figure 7.1 shows that the dots are quite scattered, it seems there is no strong relationship between students' performance on the two tests.

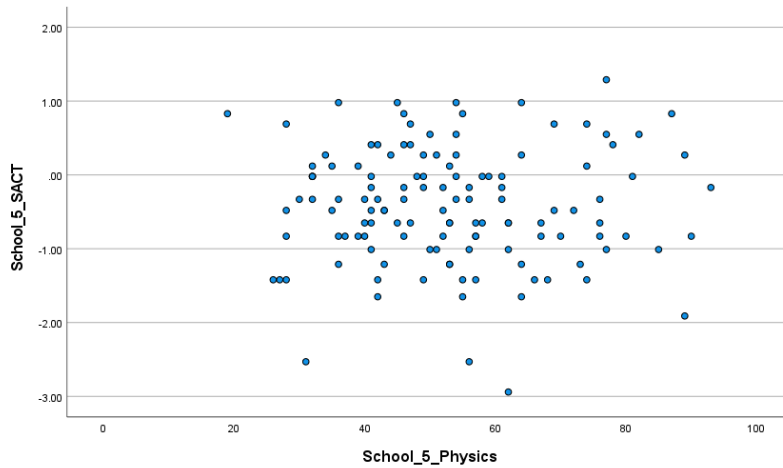


Figure 7.1 Scatterplot of school Physics scores in school 5

In terms of the distribution of the two variables, the histograms below show that scientific argumentation test (SACT) score is close to a normal distribution, but the school achievement test score is a bit skewed. Results of Kolmogorov-Smirnov test for the SACT score is  $D(126) = .065, p = .200$  and for the school test score is  $D(126) = .075, p = .080$ . Thus, both hypotheses cannot be rejected, and it can be assumed that the data are normally distributed, thus Pearson's correlation coefficient test is appropriate to use.

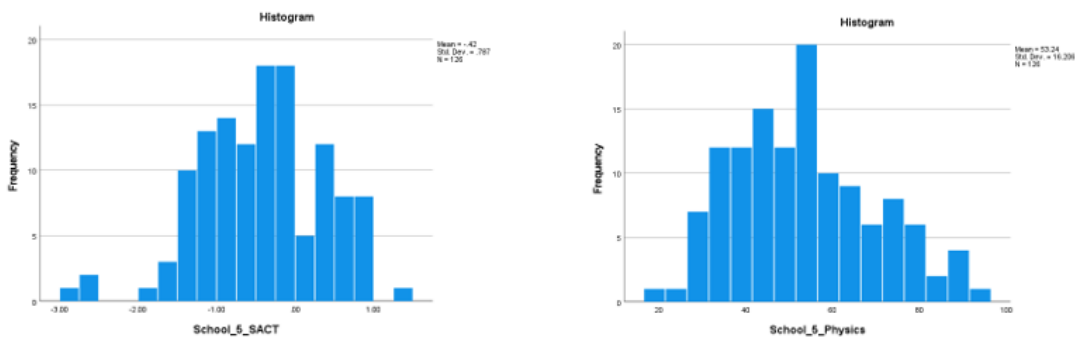


Figure 7.2 Distribution of SACT scores and school Physics scores in school 5

The results of Pearson's correlation coefficient test indicate a non-significant relationship between the school test score and SACT score with  $r = -.005$  and  $p = .958$  far bigger than .05. So, there seems no relationship between the students' performance on the two kinds of tests.

#### 7.4.2 School 6

The scatterplot below shows a slight positive relationship between the two variables.

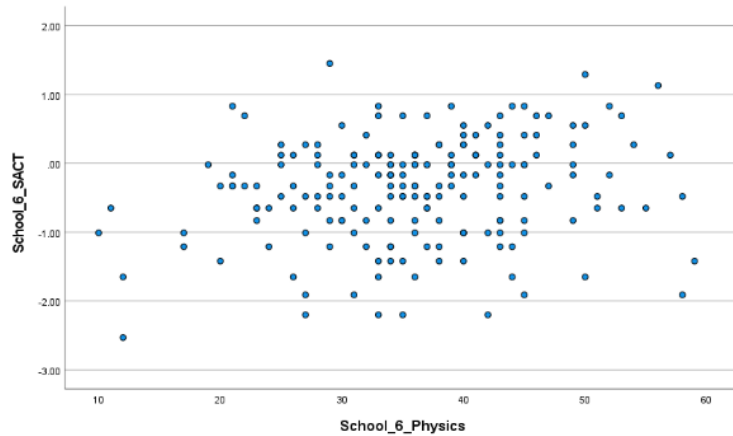


Figure 7.3 Scatterplot of SACT and school Physics scores in school 6

Figure 7.4 shows that the school achievement test score distributes more symmetrical than the SACT score, and the result of Kolmogorov-Smirnov test for the SACT score is  $D(190) = .089$ ,  $p < .001$ , and for the school test score is  $D(190) = .053$ ,  $p = .200$ . Thus, Spearman's rank correlation is appropriate to use for exploring the correlation between the two variables.

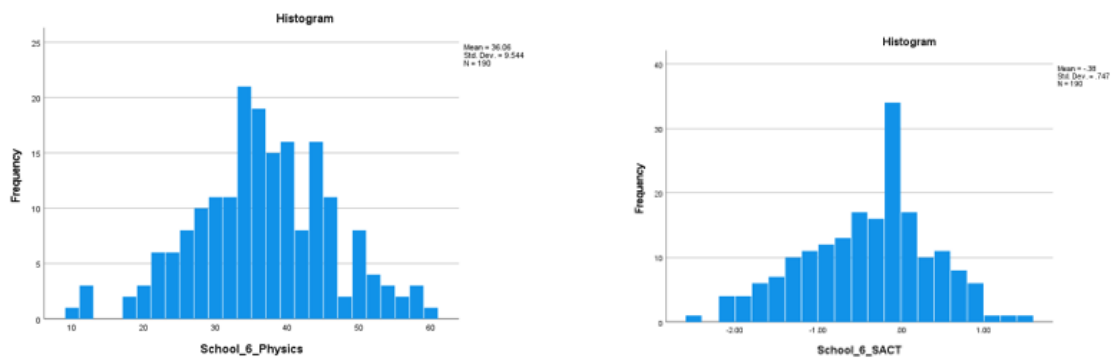


Figure 7.4 Distribution of school Physics scores and SACT scores in school 6

The correlation test results indicate that there is a positive relationship between the students' SACT scores and school test scores,  $r = .191$ , a relationship of small to moderate size (according to Cohen's rules) (Cohen, 1992), and this relationship is statistically significant since  $p = .008$ .  $R^2 = .04$ , so 4% of the variance in SACT scores is accounted for by the scores in the school Physics test, leaving 96% contributed by other factors. The relationship is still quite weak given the small value of  $r$ .

### 7.4.3 School 3

Figure 7.5 shows a more possible positive relationship between the SACT score and the school Chinese score, although neither plot shows much relationship.

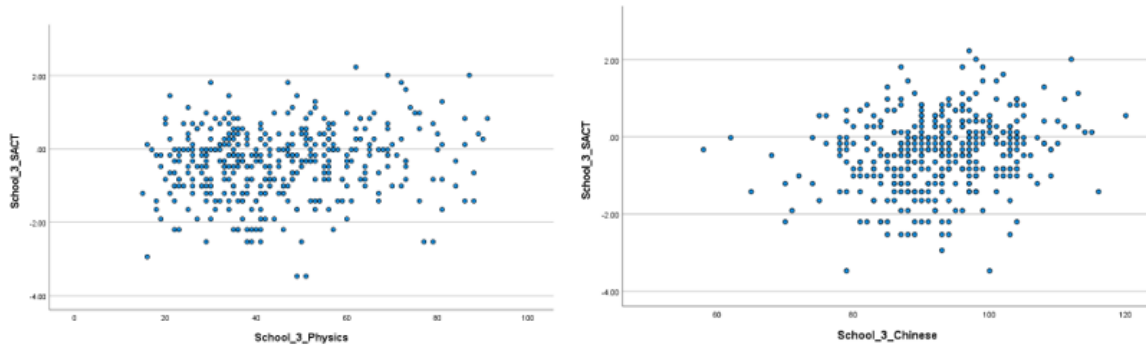


Figure 7.5 Scatterplot of SACT and school Physics and Chinese scores in school 3

The histograms below show that the SACT scores and the school Chinese scores are probably normally distributed while the school Physics scores looks skewed. The result of Kolmogorov-Smirnov test for SACT scores is  $D(393) = .081, p < .001$ , for school Physics scores is  $D(393) = .088, p < .001$ , and for school Chinese scores is  $D(393) = .044, p = .063$ . Thus, school Chinese scores are normally distributed while the other two variables are not. Spearman's rank correlation therefore is used to test the correlation between the SACT score and school Physics score as well as SACT score and school Chinese score.

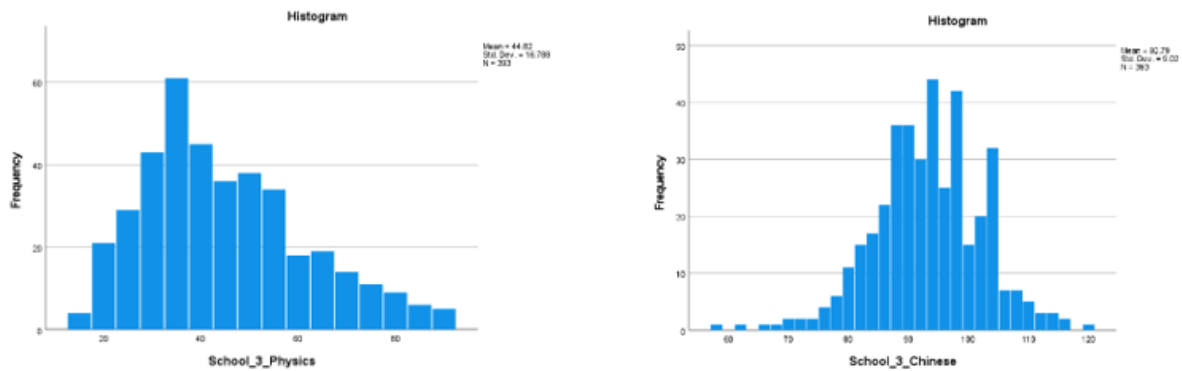


Figure 7.6 Distribution of school Physics scores and school Chinese scores in school 3

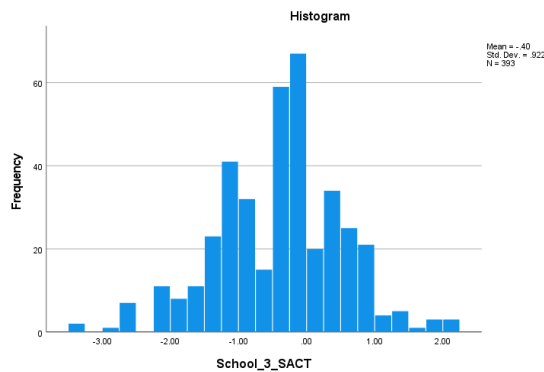


Figure 7.7 Distribution of SACT scores in school 3

The correlation test results show that the SACT score is positively correlated with the test scores from the other two school tests, and both have a relationship of small to moderate size (according to Cohen's rules). Firstly, the positive relationship between SACT score and school Physics score is statistically significant since  $p = .009$ . The correlation  $r = .132$ , and  $R^2 = .017$ . So, 1.7% of the variance in the SACT score is accounted for by the score in the school Physics test, leaving 98.3% contributed by other factors.

In terms of the relationship between the SACT score and school Chinese score, the positive relationship between them is statistically significant since  $p < .001$ . The correlation  $r = .220$ , and  $R^2 = .048$ . So, 4.8% of the variance in SACT scores is accounted for by the scores in the school Physics test, leaving 95.2% contributed by other factors. It seems that a student's SACT performance correlates more with his/her performance on the school Chinese test, although both school tests have only a small relationship with the SACT test.

Overall, the students' SAC scores seem to have rather weak relationships with their school achievement tests but where present it appears that students with stronger content knowledge (whether it be Physics or Chinese) on average do better at the SAC test. The fact that the correlations were small may be due to the fact that the content knowledge tested in the school Physics test did not match exactly with what was tested in the SAC test, or due to the fact that content knowledge plays a limited role in supporting students' SA engagement. However, the students' SAC score had stronger relationship with their school Chinese score, which may be due to the nature of SA being a discourse and finishing the SAC test needs certain written or reading ability.

### **Chapter summary**

This chapter answers RQ 4 by exploring how the context in which the students were located, the student's gender, the scaffold that was provided on the test paper, and the students' content knowledge proficiency influenced their SAC performance. Overall, students in the high-achieving school have significant better performance than students in other schools; similarly, the students in the key classes performed significantly better than the students in ordinary classes. Different from the traditional view of Chinese society on gender differences in learning science, there was no difference in SAC performance between girls and boys. However, this conclusion may not be generalized to the whole population of each gender group given not all the students in each class participated the study and the students in this study were those who

selected Physics or Science. In contrast to what the interview revealed, the scaffold provided didn't help the students to do better on the SAC assessment. Not surprisingly, students who got higher school Physics/Chinese scores tended to perform better on the SAC test, but the relationships were weak especially for that of the school Physics test scores and the SAC scores.

By exploring the SAC performance of different groups of the students, this chapter informs further understanding about the students' SAC therefore shedding light on the ways to equip students with SAC (which will be discussed further in Chapter 9).

## **Chapter 8. Understanding Students' SAC from Their Perspective**

### **Introduction**

To understand the SAC construct and Chinese high school students' SAC, Chapter 5 and 6 have presented the findings in terms of how the SAC construct has been understood and examined by its assessment and the group of Chinese high school students' SAC performance on the assessment, and Chapter 7 has explored how several SAC relevant factors influenced the students' performance on the assessment. This chapter aims to further investigate the students' SAC from their own perspectives, which answers for the most part of RQ 5. To do this, the students' viewpoints on SA, their experiences of engaging in the assessment, and their current school learning experiences are explored. As mentioned in section 4.5.2, codes were generated, and themes were constructed and modified to best capture the students' engagement in SA. Specifically, section 8.1 will introduce the participants that are involved in this chapter. Section 8.2 to 8.4 will then elaborate the three themes constructed from the interview data. The first theme is "*Students' perceptions about SA*". This theme maps what the students had already known about SA and their attitude toward SA and the SAC assessment. The second theme is "*Students benefit from taking the SAC assessment*". This theme focuses on how engaging in the SAC assessment influenced the students' understanding on SA and science learning. The third theme is "*Challenges of engaging in SA*". This theme reveals the current situation of SA implementation in school education from the students' perspective and emphasizes the difficulties they encountered when engaging in SA.

By elaborating these themes, this chapter provides more information for understanding the current situation of the students' SAC and informs the possibilities of facilitating their SAC.

### **8.1 Participants**

As mentioned in section 4.4 and 4.5, the interview data collected in the first three pilot studies were analyzed to improve the assessment, whereas the follow-up interview in the fourth pilot and the main study focuses on exploring the participants' experiences and perspectives of SA. This section briefly introduces the participants that will be mentioned in this chapter. As shown in Table 8.1, in the fourth pilot study, students (N= 18) from three cities of the same province in the middle of China were invited to the interview. There were 12 boys and 6 girls, 3 students got a score below the average score (i.e., 27) and 15 above, and 3 students took the test without scaffolding with 15 with scaffolding. Students from 5 schools in two provinces, one (N=6) of



which is in the north of China and the other (N=6) is in the south of China, were invited for the main study. There were 5 girls, 4 students got a score below the mean (i.e., 15) and 5 students took the test without scaffolding out of the 12 participants.

The sample size in the main study is smaller than in the pilot because the main data collection was administered in December when schools were preparing for the semester examination, and students had less holiday (usually one day's break every two weeks) and couldn't spare time to participate in the interview. Although most of them were still willing to attend the interview when they had the one day's break, it had been a long time since they took the test, and they would have forgotten most of it. Some students didn't have a computer or phone for the interview despite volunteering (the phone number they left on the test paper belonged to their parents); some students were not allowed to use the phone for a long time even in holiday and they easily missed my invitation message, and we had missed the best interview time when they finally had a chance to respond to me. Besides, different schools had different holiday plans and students in some schools had less holiday than others. Therefore, there were fewer participants in some schools than others. Although more students who participated in the interview had higher-than-average scores, the invitation to the participants was sent to students across different levels of performance (see section 4.3.3). The final participants in both studies depended on the above factors and was a result of tradeoff, but the results from analyzing the interview data may therefore be biased by the fact that most participants get above-average scores on the SAC assessment.

Table 8.1 Interview sampling

Fourth pilot					Main study				
ID	Gender	School	Scaffold	Score	ID	Gender	School	Scaffold	Score
F1	Boy	2	Y	25	M1	Boy	2	Y	33
F2	Boy	3	Y	42	M2	Girl	1	Y	27
F3	Boy	4	Y	33	M3	Boy	5	N	26
F4	Boy	2	Y	29	M4	Girl	3	N	15
F5	Girl	1	Y	20	M5	Boy	6	Y	12
F6	Boy	1	Y	35	M6	Boy	3	Y	11
F7	Girl	2	Y	38	M7	Boy	1	N	22
F8	Boy	1	Y	42	M8	Boy	1	Y	21
F9	Girl	2	Y	35	M9	Girl	5	Y	20
F10	Boy	2	N	31	M10	Girl	2	Y	33
F11	Girl	3	Y	31	M11	Girl	1	N	28
F12	Boy	3	Y	47	M12	Boy	3	N	9
F13	Boy	2	Y	30					
F14	Boy	1	N	33					
F15	Girl	1	Y	25					
F16	Boy	3	Y	38					
F17	Girl	1	N	51					
F18	Boy	4	Y	31					

## 8.2 Students' perceptions about SA

Two themes are constructed in this section to show the participants' perceptions on SA, namely 'Existing awareness of SA transferred from previous experience' and 'Positive attitude on SA and the assessment'. It is of value to know what students already knew and how they think about SA in general and in the context of school education because students' perceptions on SA affects their engagement/willingness of engaging in it (Kuhn et al., 2010). Each of the themes will be explored in detail subsequently.

### 8.2.1 Existing awareness of SA transferred from previous experience

Theme 1 starts by exploring how much the participants have already known about SA and how they got that understanding. It is important because knowing where they are now can inform how they should be taught next. High school students of their age already have some life experience, and the data shows that they had had a certain level of understanding about SA by referring to their previous life and study experiences.

Most of the participants (28 out of 30) said that they had "never heard of" (F18 etc.) SA previously but their learning experience in other subjects helped them understand what's going on in the assessment, such as M12 who mentioned that "it's quite like the argument essays task in Chinese...it requires claim and evidence as well".

More in-depth, previous life experience made students realize that they had “experienced and seen others present similar thinking process before” (F12) as they engaged more with the assessment tasks, although it was their “first time to know the terminology of SA” (F12). M10 further talked about her experience when she mentioned that “I am not unfamiliar with the behavior (of SA) while never heard of this term”, as shown in the following excerpt:

“I actually experienced it a lot in life, for example, whenever I saw a problem, a claim would jump out of my mind, and I would become more and more believed in it through my reasoning process. Then one day, something else might remind me that my previous claim might not be perfect and there would be conflict between the previous claim and the new claim. A more comprehensive conclusion would be constructed by the communication of the two claims...I like this activity for it can make my own claim more comprehensive and rigorous, by interacting with others who hold different claims.”

There were six other students who talked about their previous experience of doing the activity or watching relevant TV shows and books, six out of the seven students (M2, M1, M8, M10, F12, and F16) got relatively higher scores on the SAC test. These six students either to some extent had trained themselves to possess this ability explicitly or attended in debate competition before, and they showed deeper thinking in what SA is and how it had pervaded their own life and study. Although their experiences had not been from the field of science, the pattern of their performance echoes with Groom et al. 's (2018) finding that the familiarity with the epistemic characteristics of SA contribute to the ability to engage in the activity. Therefore, although SA is a field dependent activity in which scientific knowledge and the understanding of the nature of science play important roles, the finding indicates that understanding or skills of argumentation obtained from other fields may transfer to argumentation in science. This is in accordance with Bricker and Bell (2012) that discuss how students' everyday argumentative practices should be appreciated and be taken as a leverage for argumentation in school science learning.

Students tended to understand SA from a more cognitive aspect compared to social aspect, such as “logical thinking” (F2), “using evidence to prove hypothesis” (F3), “analysis of a problem” (F5). Although several students understood SA as understanding others thus gaining deeper understanding on the topic that is discussed. Most students' understanding of the social aspects of SA emphasized the competitive rather than the collaborative aspect. That is, students tended to view the social aspect of SA as people aim to persuade others or win the conversation in which there are different claims rather than to construct the conversation collaboratively to reach consensus. Additionally, a few of them implicitly mentioned, “I am not that kind of

person that like to argue with others, it is their right to have their own opinions” (M7) and “some students are born aggressive, and they always like to challenge others” (M6). Both these students talked with an emotional tone that suggests a negative signal about arguing with others, but they tended to refer to everyday life examples when saying that and became appreciate the activity when talking about it based on what were demonstrated in the assessment.

Therefore, students actually had known that SA is a socially relevant activity although their understanding had been less comprehensive and depended on contexts. As revealed by Bricker and Bell (2012), middle school students tended to associate the word ‘argument’ with behaviors such as yelling and fighting when they talked about the word in general, but their views became broader and more nuanced when talking about it in a specific context of argumentative activity, such as associated the word with discussion and decision-making. In this study, the reason of the students’ talking about SA from a more cognitive perspective is probably because they were talking about SA with the influence of school learning where analytical forms of thinking are appreciated. Their social perspective mainly transferred from their life experience where arguing is to persuade others and win the competition. In addition, as McNeil (2011) found that 5<sup>th</sup> grade students tend to understand argumentation as “emotional or angry fighting” (p. 801), some of the participants conveyed similar understanding of SA when it came to its social aspect. These imply that school education was some way short of creating an environment of constructing knowledge socially and especially collaboratively.

Students shared their understanding about what high-quality SA was as well, using phrases such as “multifaced and sufficient evidence” (M2), “logical coherent” (M10), “includes the four elements” (M8), “closely related with claim” (M1), and “persuasive” (M7). Most of these participants didn’t demonstrate this understanding of high-quality arguments to their evaluation/production of SA, but these understandings revealed that students should be able to form a more complex understanding about SA when given appropriate support and opportunity to engage in it. In terms of the excluding characteristics of SA in Physics, some of them mentioned that facts or formulas should be used to prove the hypothesis. For example, M7 said, “Probably, it’s similar to proposing a hypothesis about a Physics phenomenon, and to explain it using facts”.

The above statements show that the participants generally already had an existing understanding of SA either by understanding its literal meaning or through the transference of their previous experience, while the assessment provided them with an opportunity to explicitly

think about it. Their understanding was prone more to take SA as analytical and logical rather than dialectical and rhetorical, and much less to pay attention to the nature of the scientific enterprise. This implies that argumentation had been happening everywhere in the students' life, but no one had shown it to them deliberately especially in the context of science education, which suggests that the students had never been taught about SA explicitly. Students' experience to some extent shapes their ways of understanding the world, similarly, learning happens through the transformation of experience (Kolb, 1984). Although students' understanding may in part be influenced by the presentation of the SAC assessment (i.e., does not convey much about the dynamic social aspects of SA), it seems lacking opportunities in the current school science education for students to experience the social process of SA.

### **8.2.2 Positive attitude on SA and the assessment**

This subtheme explores whether the participants think it is worth engaging in SA by examining their attitude to the assessment and to SA in general. After taking the assessment, most of the participants (18/30) talked about their positive experience explicitly, using phrases such as that they found the items “interesting”, or that they “enjoyed” doing it. The remaining participants shared their feeling of unfamiliarity and how different it was compared with their normal tests, and they didn't show a sign of any negative experience despite some of them feeling it “difficult”. One thing to note is that from the interview, whether students found the assessment interesting was not necessarily related to their scores. Students who got lower scores also expressed their experience in an explicitly positive way. Therefore, the positive experience of participants is not necessarily attributed to their outstanding performance in the assessment.

The feelings of “interesting” and “difficult” seem to have the same roots. Because “unfamiliar” material made the assessment “difficult”, as stated by F17 who mentioned “the difficulty lies in that it is a new thing to me, and I am totally unfamiliar with it”; while the “novel” material made it “interesting”, as F16 said:

“These are not accessible in our usual study, and I felt it's quite novel...we didn't have this kind of discussion in school...I felt excited when I got this test, it's like finally I got a chance to present my, uh, to try this stuff.”

However, students' positive experiences were not merely because of their curiosity about new things, they mentioned their preference on thinking about problems that are close to real life. Over 20 students mentioned that some scenarios in the assessment were close to daily life and 11 students explicitly expressed that they “like” thinking about real life problems. Such as F10

who said:

“Most of these scenarios are quite close to life, such as...sometimes I also think about these questions when I saw these phenomena, but seldom pay attention to it...it’s quite interesting to think over these questions.”

This corresponds with the long-standing call for learning and assessing to have authentic context (Archbald & Newman, 1988) and the awareness that learning and performance depends on context and motivation (Wiggins, 1993). Real life problems motivate students’ desire of exploration and give them the sense of accomplishment. As pointed out by Cumming and Maxwell (1999), students’ learning outcome can be enhanced if they perceive the relevance of learning or assessment activities. As M12 illustrated:

“I really think we come to school to solve more (problems in real life), to know more scientific knowledge that happens in our life...for example, if something broke in my house, like electrical stuff, I can help my parents to fix it rather than to repeat the examination items once and once again on the text book, some of them, you know, too far away from our life and it’s just hard to imagine them.”

Another significant feeling of participants is that most students (N=26) mentioned that SACT assessed more about their thinking process, and many of them expressed their endorsement of this experience. Students found they were “actually thinking about the scenario and the problem while doing SAC tasks” (M3), while they “only focused on providing the right answer when doing normal test items” (M3). F3 provided a detailed description that:

“Unlike our normal examinations that assess knowledge and ask us to figure out the final answer follow the certain way we have been taught, this test is closer to real life and more practical, and cares more about the logic and thinking ability...like task 4, it provides several information and we need to propose and justify our claim based on these information, which is quite flexible and we have to pay attention to our own thinking...what is assessed here is like a whole set of stuff that is quite systematic, and I found it particularly enjoyable anyway.”

On the other hand, this characteristic of SACT made some students feel it was more difficult compared to their normal test. M12, for instance, compared the difference between SACT and their usual tests as:

“It is even easier for me to throw several formulas than to justify these questions, at least I know some relevant formulas when doing our usual tests, but now, it is hard to think about how to justify.”

As Baird et al. (2017) emphasize that the educational concerns should be the priority of educational assessment. If an assessment makes ‘throw in relevant formulas’ to be the first

reaction of students, the assessment is very likely to end up short of serving its educational objectives. Providing students with the opportunity to think about a question and to gain potential new knowledge should be better than to go directly to the application of formulas without really thinking about the problem. Given SA is now specified in the Physics curriculum, the current school assessments appeared, from the students' perspectives, to be short of assessing SA. Despite the positive feedback from the participants, the quality of this assessment should not be exaggerated for there was one participant mentioned that "the thinking process is still a little bit short, and it is not sufficient to enjoy the fun of thinking, since the connection between claim and evidence is quite direct" (M10). So, as an exploration, the assessment was appreciated by the participants while there is still room for improvement, which suggests that the participants held a relatively positive attitude toward the assessment and thereby the construct it was assessing (i.e., SAC).

In terms of SA in general, it seems that high school students at their age had reached the cognitive level to recognize the role SA plays and its value in their life and study. A certain number of participants (N=11) who were asked (N=12) talked about whether and why they thought SA was valuable for their study and daily life, and they had their own thoughts on what they expected learning should be like. However, most of them (N=8) did not go into detail about why they thought it was valuable to engage in SA, they tended to say something general and superficial such as "it is useful to help us obtain logical thinking" (M2), and "it helps with our thinking ability and to learn knowledge in more depth" (M11). As mentioned by Schwarz and Baker (2016) that people's values towards argumentation have an important effect on their performance, students who expressed a higher degree of recognition for the value of SA tended to engage better in the assessment and talk more in the interview, for example:

"Argumentation is everywhere, we always come across problems that need to be solved by argumentation in life and in study... if we want to really understand what we learned and be able to apply it, we have to have this kind of thinking ability...and another reason why I like it is that I want to enhance or improve the reasonability of my own argument by arguing with others, because it can make me reflect on my own thinking and reveal the problems in my thinking...learning should not be repeating the examination items, and high score doesn't mean high logical thinking ability...we need something to really think over." (M10)

But when the participants situated SA into the current education context, they became pessimistic in terms of the 'usefulness' of SA. Such as M12 said "yeah, it is worthwhile in itself, but it's hard to say given the situation we are in, you know, if it cannot help us gain

higher examination scores”. Similar viewpoints prevailed in the students’ interviews, as shown in the following excerpts:

“I quite like arguing or discussing with others in life, but not very sure how it can be conducted in school learning. We students mainly focus on doing examination papers and getting high scores, so not sure that students would like to spend their limited time on engaging in SA, you know, it is not included in examinations.” (M6)

Likewise, as mentioned in section 4.6, students had the right to not submit/respond to the test paper, the participants who talked about their classmates who didn’t take the test mentioned that it was because they found the test difficult or they thought it was useless to do the test.

Overall, students generally held a positive attitude toward the test and SA itself, which justified the value of this assessment and shed light on the kind of assessment that should be advocated, although students’ attitude tends to be twofold: theoretically optimistic and practically pessimistic. This is in accordance with Liu and Helwig (2020) that Chinese students view the aim of education from its intrinsic value while view the school education they received (or will receive) as having more extrinsic value (see section 2.3). However, one needs to be cautious when interpreting the results of this study considering most participants got above-average SAC scores.

Considering that students are supposed to benefit from teaching, learning and assessment, it is significant to listen to the voice of students/test takers as it is often overlooked (Butler et al., 2021; Elwood et al., 2017). From the perspective of learning rather than the external need of obtaining high scores on high-stakes examinations, what the students expect for their learning is aligned with the long-standing aim of education proposed by Dewey such as fostering students’ progressive and life-long development of active thinking (Kohlberg & Mayer, 1972). Thus, integrating SA into school education meets the students’ expectations for learning.

### **8.3 Students benefit from taking the SAC assessment**

This section investigates the participants’ experience of engaging in the assessment with a focus on what the assessment brought to them. Probing into what engaging in SAC assessment brought to students can help us figure out the educational value of the assessment and thus inform the kind of assessments that may benefit students’ learning, and further understand what is overlooked in the current school education. Two themes, namely ‘Pedagogical function of the assessment’ and ‘Introspections through the assessment’ emerged to illustrate the



participants' interaction with the assessment.

### **8.3.1 Pedagogical function of the assessment**

All the students talked about the new insights or knowledge they gained from the test except for two participants who said that they didn't feel they learned anything new by engaging in SACT. For those students who did the test paper with scaffold, they talked more about they "became clear about the meaning of the four elements" (F9) described in the scaffold, which is quite straightforward and unsurprising. But in practice, they still tended to provide an overall explanation just after their claim to explain why they held that claim not the other claim, rather than thinking over the evidence, reasons, and rebuttals in their argument. This resonates with Berland and Reiser's (2009) study, which found that middle school students tended to weave these elements together in their argument rather than articulating each of them. Thus, it seems there is a gap between what they knew and how they performed, and the scaffold played a limited role.

In addition to learning from the direct information provided for them on the test paper, many students (N=15) talked about their awareness of "the process of scientific argumentation" (F17) and "how the activity worked as a system" (F14) as they progressed through the test. As M10 said:

"At the beginning I was surprised to do the test, after I saw the definition of the four elements, I started to think...when I was doing the first two tasks I was still confused, then I continued to understand it, to feel it, when I proceeded to the last two tasks, I felt I was almost there...the whole test gave me a feeling of, like a system, I started to have some understanding about the process of SA."

One thing to note is that the students who did the test paper with scaffold tended to use and emphasize the four elements in the scaffold across the interview, while those who did the test without scaffold explained their understanding using more normal language such as 'logic' and 'debate'. Despite students who did the test without scaffolding not expressing themselves using the terms, their understanding was not necessarily poorer than those who had scaffolding, such as M3 who got a high score in SACT and showed relatively deep understanding of SA during the interview. This echoes what have described in section 8.2.1 that the students had a sense of the activity while were unfamiliar with the terms. But the scaffolding indeed supported them to think explicitly using Toulmin's terminology.

Some of the participants (N=6) also talked about how the assessment aroused discussion and

argumentation among students. They found it “fascinating” (F7) that people hold different claims on an issue. This is consistent with the finding in section 8.2.2 that the students valued the role of SA in learning. F11 provided a more vivid description of the situation of their discussion:

“I remember the next lesson was a seminar, and several of us didn’t listen to the seminar but were discussing about the items. We were so happy when we were discussing since everyone shared their thoughts which were different and several students from other classes seated near us also joined our discussion...After this experience, I found that it is interesting when we discuss a problem with uncertain answers and close to life, and we also found it is important to think about the problem from other’s side and try to understand what they mean.”

These data show that the assessment functioned pedagogically in general, but as revealed in section 8.2.1, students had different levels of existing understanding about SA, so the assessment brought different degrees of knowledge to different students. The finding also reveals that almost all the students had never encountered these terms in the context of science study, and never organized their thinking using the four elements of SA intentionally. Due to the limitation of pencil and paper tests, the social aspect of SA was presented in this assessment in a rather ‘fake’ and implicit way, but this design indeed aroused students’ awareness of the social part of SA. Overall, the assessment design that tried to systematically represent SAC - an unfamiliar concept for the students- made the students aware of SA and they learned more about SA from the assessment. Moreover, assessments that value educational objectives, in this case scientific argumentation, can contribute to students’ knowledge and enthusiasm for learning.

### **8.3.2 Introspections through the assessment**

Students reflected on their understanding about SA by comparing what they had already known with what they learned from the assessment, which shows that they had some awareness of what they were doing and that contributed to their knowledge of SA as well.

Either by reflecting on their performance on the assessment or their understanding of SA, 11 students mentioned that they needed to “improve (their) expression skills to make the argument more succinct and clearly organized” (M1) and they realized that SA requires people to “organize their language in a good way” (F2). Having seldom been engaged in activities that need them to propose their own claim, students (N=6) mentioned that they didn’t realize before the importance of putting forward and justifying their own claim when “people have different

opinions on a problem” (F18). Meanwhile, having always been involved in providing answers without ‘co-operation’ or ‘argumentation’ and “trying to prove that my answer is correct” (F14), 8 students mentioned that they started to be conscious of “listening to others’ viewpoints” (F14) and “understanding others’ opinions” (F11). This is consistent with what was found in Yun and Kim (2015) that students tend to ignore what others say and focus on their own viewpoints. This phenomenon indicates that students had lacked the intention of attending to others’ positions, which is at the heart of argumentation (Kuhn & Udell, 2007), and lacked the sense of communicating and engaging in conversation which is essential for knowledge construction (Lemke, 1990). In addition, it is worth noting here that all the three reflections mentioned above are related with the social aspects of SA, which supports the finding in section 8.2.1 that the students had fallen short of understanding SA from a social perspective.

Moreover, students mentioned that they had “never thought *reason* is used to connect between claim and evidence” (F16) given it is a widely used word in daily life that indicates “to explain the cause of a behavior” (F16). They were aware of using *reason* to construct the relationship between claim and evidence to make their claim “rigorous” (M3) during the assessment. Students’ understanding of *reason* was prone to be in the category of ‘conventions’ and ‘stories’ according to Tilly’s (2006) categorization of reason, which are more about commonplace reasons and explanatory reasons. While both Toulmin’s argumentative account and science practice appreciates *reason* as ‘code’ and ‘technical account’ that care about causal accounts and reside in certain professional fields (Bricker & Bell, 2012). Students’ nebulous understanding on the connection between evidence and reason is also a manifestation of the lack of training on argumentation especially that uses Toulmin’s framework.

Although students gained some new understanding about SA, the pedagogical influence of the assessment was limited and probably not long-lasting. This can be uncovered from the students (N=7) inconsistent performance on SACT which revealed their unstable understanding about SA. There were two manifestations of the inconsistencies, the first was that they provided quite good understanding theoretically but performed differently in practice. The second was that they showed contradictory understanding of the same SA element on different items. For example, F6 explained his understanding on evidence and reason quite well in that “now I have a very clear understanding toward reason and evidence: evidence is the fact exists objectively, and reason is used to explain why the evidence supports the claim”. But when he was asked to explain how he figured out Ie\_3.2 (Identification of Evidence item, see Appendix 19) he further

explained that “well, he said that ‘according to facts b and c, he felt the black ball falls faster’, you know, he used the word ‘felt’, which is quite subjective. This should not be objective evidence that can justify himself”. He ignored the evidence (facts a and b) in the argument and mistook the reason as evidence.

Another example is M8’s understanding on evidence and reason on items Ee\_3.1 (Evaluation of Evidence), Ie\_3.2 (Identification of Evidence) and Ir\_7.1 (Identification of Reason) (see Appendix 21). For Ee\_3.1, he said that the provided evidence was ‘relevant’ because “the fact is a characteristic that the ball possesses even if the evidence might not support the claim”. But in Ie\_3.2, he said that “‘facts b and c’ are not evidence, because these two facts cannot support the claim”. Simply speaking, in Ee\_3.1 he thought that the fact that objectively exists and is a physical property is relevant evidence even if it cannot support the claim, but in Ie\_3.2 he thought that facts that cannot support the claim are not even evidence of the argument. Further, in Ir\_7.1 he took fact a, which is also a physical property, as a reason. However, when he was asked about his understanding on evidence and reason, he said that “evidence exists objectively, and it is real and doesn’t need extra decoration or explanation. But reason is what helps to construct an argument by observing or analyzing the evidence”, which is a quite good understanding.

The possible reason behind the inconsistency could be because of the scaffold provided in the test. Students gained their understanding by learning timely and temporally from the scaffold or by recalling their previous experience, but this understanding was not solid and far-reaching enough for them to apply it in practice. This is consistent with the findings in section 7.3 that there is no significant difference in SAC performance between students who were provided with the scaffold and those who were not.

Overall, there seems no doubt that the assessment brought students some new understanding of SA, and they even applied some of it immediately in the assessment. It can also be assumed that the SACT had an educational value for most of the participants, as they were thinking and reflecting when they went through the test, as described by Dewey (1949) as ‘learning from experience’. These new insights the students got from the assessment informs what has been overlooked in school education in terms of equipping students with SAC. Nevertheless, their understanding and awareness about SA seemed unstable and superficial, and there was a gap between what they knew and whether they could apply it in practice. Therefore, if there is no follow-up teaching and practice, the impact of the assessment would be short-lived for the

students, and they may quickly forget these gains.

## **8.4 Challenges of engaging in SA**

This section focuses on what may have impeded the students' engagement in SA. Understanding the challenges students face can help facilitate its teaching and learning. Two themes, namely 'Lack of opportunities to engage in SA' and 'Difficulties of engaging in SA' were generated and will be elaborated upon in detail.

### **8.4.1 Lack of opportunities to engage in SA**

The participants were despondent when talking about their current experience of learning and taking assessments in schools. Some students realized the fact that they seldom engaged in SA despite its value is probably because "it is not included in the school examination and college entrance examination" (M3). In addition, "the teaching schedule is too tight" (M9) to support this activity. They expressed that the "repeated" (M12) practices on examination papers gave them the feeling that:

"The examination items we usually took, and what we learned in the classroom, it's like using formulas to get the answer... In most cases, these learning (take examinations) experience was like enabling us with muscle memory and the ability to control symbols, they did not make connection with real life or with logic." (M10)

Corresponding to their comparison of SACT and their normal tests, they pointed out that the current assessments "care more about the final answer that is derived through applying a combination of formulas being correct" (F4). Given the significant position of the high-stakes examination of the Gaokao, it goes without saying that ensuring students get high score in the Gaokao is the major priority for high schools. Students felt hopeless in terms of integrating SA into the current classrooms, such as M6 who said:

"You know what school and the society is expecting of us? High scores. We seldom have holiday because of this. We are expected to get higher scores by study, or more accurately, doing examination papers repeatedly... as for SA, there is no time for school and teachers to give it a concern."

They shared their concern on the contradiction between the current education system and promoting SAC as well, for example M10 mentioned:

"There is one interesting phenomenon of school education, its goal is to gain higher scores, but you know, there is no direct relationship between the scores you get and your argumentation ability."

Assessment tells teachers and students what is important in learning, thereby influencing students' ways of learning and thinking (Cumming & Maxwell, 1999; Baird et al., 2017). Some of the participants had a clear awareness of this. F15 expressed that she "felt good to go out of the zone" in which their "thinking has been to some extent fixed in a pattern". In terms of the concrete manifestation of the "fixed mindset" (M10), M10 gave a detailed discussion about this:

"It is horrible that many students at our age have already had very strong mindsets...there are many representations of people's mindset, like some of us are afraid of going outside the comfort zone to be creative or to solve problems we never met before. Another example is that we started to do examination papers from an early age, and we have done like hundreds and thousands of test papers, and all the questions have one answer, either right or wrong, when we met a question has a second answer, we are afraid to choose it even if we got it through reasoning."

Some other clues about this "fixed mindset" can also be found from their performance on the assessment. Such as F12 said:

"Yeah, I knew that. In fact, I thought this option might be right and I was hesitant whether to choose it. But I didn't dare to choose it for it is too absolute."

Consistent with their perception that SACT is different from what they usually do in that it pays more attention to the thinking process, answering the "why" question made some students feel like "freaking out" (F5). F12 provided a more detailed illustration on this common phenomenon:

"Many schools do not emphasis thinking, they care more about keeping doing examination items and believe in short cuts. Just follow what teachers told you, and to get scores by memorizing the formulas and conclusions that teachers have summarized for you. It is not difficult to get high score mostly by memory, teachers tell you the beginning and the ending of the story and leave apart what happened in the middle. In most cases, I think what happens in the middle is the argumentation process, and I believe most students don't know how to tell the whole story by adding the process. After all, it is much easier to memorize conclusion that to figure out how to get it."

The emphasis on the results rather than the thinking process can also be reflected from some students' way of figuring out multiple-choice items that "it always works by choosing the positive answer if you feel what this person says is plausible, and you don't have to think over it very carefully" (M12). Similarly, M6 mentioned that,

"Like we usually do when taking examinations, if you find an option that is way too absolute, then this option must be wrong. It is always safe to choose relatively neutral one especially when you are not sure about the answer, that's what our teacher taught

us.”

Accordingly, the final answer means a lot to students since they didn't think they performed well because they “didn't even know the answer for several questions” (M9), and it is also of significance to their judgement of an argument, like M8 said,

“Well, I chose the option of ‘none of the above’ because he gives the wrong answer. I didn't think too much on his argument after I found his answer is wrong, and I just choose ‘none of the above’.”

High-stakes tests always drive teachers and learners to change their behavior in a way to maximize the rewards of teachers, learners, and their institutions (Stobart & Eggen, 2012; Madaus, 1988). It is not hard to understand that a shortcut becomes a choice when students don't really understand a question while ‘getting the item right’ is an urgent need for them. Accompanied with the demand of high scores on high-stakes examinations, students actively conformed with the current rules imposed on them despite their awareness of the disadvantages. As reflected from what M12 said that “but I still have hope for life, although (what I learned) cannot be applied in life, but I still feel happy if I can do better than others...I can still have a feeling of accomplishment”. They were clear about their purpose that “I need to get through the examination anyway” (M10) although “the process is painful and uncomfortable” (F12) and they “hoped to be provided with opportunities for developing other abilities such as argumentation” (F11).

Overall, the current school assessment and teaching seem to fall short of paying attention to the students' role as learners and caring about their ability in scientific argumentation. Either actively or passively, students were involved in a cycle of ‘learning for examinations’ that does not support learning, they had no choice but to follow the rules despite their awareness of the problems in their learning. This is not consistent with what the Curriculum document expects for students. Despite high-stakes examinations such as Gaokao have been perceived as necessary for the condition of China and motivate students to study with more effort (see section 2.3), the alignment between students' own reflection, their performance on SACT and the current state of school education implies that:

- 1) the dominant focus of school education is still what students know rather than how they know what they know and why they believe what they know,
- 2) there is a lack of emphasis on SA in current school education and
- 3) the current assessment drives the teaching and learning in an undesirable way.

#### 8.4.2 Difficulties of engaging in SA

Students talked about how they figured some of the items out in the interview, which uncovered problems- they may not realize- that many of them had in SA. Firstly, the participants showed less awareness/ability to evaluate what was provided to them. Some of them (N=10) usually believed in what they saw and thought that the provided information must be “correct” (M5) especially when they were solving E-SA items and tried to explain what they saw even though they didn’t know why. As reflected in M5’s interview where he said that “maybe option B is right, but I chose C because I think that he must be able to hit the black ball using the method he described above, otherwise he wouldn’t have said that”. Similarly, there were students who thought that every piece of information “must be useful and are supposed to be used” (F8) in their argument. Other students talked about how they found it overwhelming to deal with too much information in the tasks that they needed to analyze and compare each piece of the information.

Another common problem that happened for students (N=14) is that they tended to rely on intuition rather than evidence and reason in SA. Like one of the participants who said, “my claim is that aiming at the bottom of the ball can hit it...just my intuition, a feeling” (F3). Except for the word ‘feeling’, they also liked to use the word ‘common sense’ to support their claim. But their understanding about ‘common sense’ was not exactly what it should mean, it was not like ‘the salt is salty’, but they understood it as an intuition or a feeling. As one of the participants M8 said,

“It’s common sense, a feeling...like if a person who never learned Physics, he or she would probably use their common sense to think that the heavier ball falls faster, but I learned Physics, so I know that’s wrong.”

While students realized the importance of proposing their own claim, some students found it difficult to make decisions because they either think “both sides seem reasonable and didn’t know which viewpoints should be supported” (M5) or found that they themselves “were not used to proposing their own claim and justifying it” (M3). After they made their claim and tried to justify it, they were also thinking about “the other’s argument to be reasonable” (F3). Thus, some students described themselves as being ‘hesitant’ or ‘unconfident’ in SA.

Another problem exposed in the SA process was discipline specific. Some students (N=12) either excluded “calculation using formulas” (M4) from “thinking over” (M3) and “understanding” (M4) Physics problems or felt “unwilling” (M9) to use formulae to solve an



unfamiliar Physics problem. Tasks 3 and 7 were designed to allow students to use the very basic formulae they had learnt to form their arguments. However, few of them did that although they knew the formula, which is contradictory to their description of ‘getting used to taking tests that use formulas to get conclusions’(M3). As F7 said,

“I saw the formula given in ‘fact f’, but just an impression, and I didn’t use it...I didn’t calculate, just thought the ball is under the force of gravity and should fall faster.”

‘Formula’ and ‘calculation’ seem to become the representative characteristic of the normal school test from the students’ perspectives, and they tended to exclude it from understanding Physics phenomena and solving Physics problems.

Except for the problems that emerged from the interview, students themselves also shared the difficulties they faced, or they thought they may have when engaging in SA, which are quite consistent with their reflections in section 8.3.2. For quite a few students (N=9), they felt “confused about the difference between reason and evidence” (F3), which is consistent with Berland and Reiser (2009) and Sadler (2006) that students have difficulties in distinguishing between evidence and reason. In addition, students tended to have a relatively weak understanding towards reason compared to evidence. They usually “found that it has been hard to connect between claim and evidence accurately” (F16) and they “don’t really know how to construct a coherent reason to explain the relationship between claim and evidence.” (F13). This resonates with Deng and Wang (2017) that it was more difficult for Chinese students to construct warrants than evidence.

As mentioned in section 8.3.2 the students reflected upon their unawareness of listening to others’ viewpoints, they had difficulties in “thinking from another side” (F14). As reflected from their test papers, many students were trying to prove themselves as right without mentioning why others were wrong. Such as M2 said “I just don’t know how to rebut others, I felt that my rebuttal is the same as my reason”. Students’ inability of addressing alternative claims has been identified by previous studies (Garcia & Andersen, 2007; Jiménez-Aleixandre et al., 2000), which indicates their deficient understanding about the norms of argumentation and further imply the necessity of providing students with more opportunities to participating in argumentative or cooperative activities where different voices and thinking can be exposed and shared with each other.

Parallel to their awareness of the importance of language in SA, some students (N=12) felt it

difficult to express what they thought in a “clear and accurate” (M3) way either in written or verbal form. Such as M4 who mentioned that “I feel I need to improve my expression skills, it’s like sometimes I know, but it’s just hard to express what I want to say”. Students’ difficulty in expressing themselves in the context of science learning is not a coincidence of this study, its existence can be traced back to the long history of separating knowledge into different subjects especially high school subjects in China have been organized into two categories (i.e., arts-stream and science-stream) for decades (mentioned in section 2.3). As mentioned by Yore et al. (2003), traditional pedagogical culture emphasizes abstract knowledge while viewing language activities as marginal to science learning, thus students had little awareness of the role of language in science. However, SA is a social activity that requires communication with others (Duschl & Osborne, 2002). Some students’ performance on the test paper was better than in the interview when they were asked how they figured the items out, such as M4 and F18, while some students presented better understanding in the interview compared to their test paper such as M9. Some students themselves also realized that one language form tends to be easier for them than the other, such as F12. Thus, both of students’ verbal skill and written skill need to be emphasized in school education to empower them with more possible ways of sharing their voice with others.

So far, I have talked about a few points for being involved in SA of which the participants realized the importance of SA and they encountered difficulties in performing SA. We can see that being aware of the importance of a skill does not equate to being able to apply the skill in practice. Detailed instruction and practice are needed to enable students to transform what they are aware of to what they understand, and to transform both to applying them in practice.

Moreover, many students mentioned that psychological quality is an important factor as well especially if they needed to argue in front of other people. Such as M1 who said, “if face to face, I would be very nervous, I would forget what I was planning to say”. The question of how big a role participants’ psychological situation plays in engaging in SA was mentioned in previous studies as well. As suggested by Russell and Aydeniz (2013), peer pressure influences students’ performance on SA tasks by making them reticent. Students who self-reported with low school Physics test achievement tended to be less confident and talked less in the interview. Likewise, relatively higher levels of prior knowledge should have made students feel safer when dealing with the tasks because they were more confident at expressing themselves. Similar than their focus on the correct answer while contrary to their appreciation of SA, some

students shared that “it would not be good if I said something wrong (in the classroom) while other students are correct, I would feel quite bad” (M12). Thus, when talking about the competence of engaging in SA, the behavior presented by students may have limited indication. Except for what the students are ‘able to do’, it is also important to understand what influences their ‘willingness/motivation to do so’.

Almost all the participants mentioned that sufficient knowledge about SA and content knowledge is important for being involved in SA. Such as F5 who mentioned that “I felt I would have more to say and know what to say if I was familiar with the content and context of the task...and I will perform better if I know more about SA”. Students who mentioned that they were familiar with the content in a task tended to show more confidence and to talk more about that task, and the quality of their argument tended to be higher compared with their performance on other tasks. This finding agrees with other studies regarding the influencing factors of SA engagement (Von Aufschnaiter et al., 2008) where prior knowledge is highly related to students’ argumentation performance and previous studies also indicated students’ awareness of the activity plays an important role in argumentation performance (Nussbaum et al., 2008; Duschl & Osborne, 2002).

Meanwhile, although we have no idea how the students who read the scaffolding would perform on the assessment if they were not provided with the scaffold and how the students would perform if their content knowledge were more proficient, students’ understanding improved after discussing with me about either the content knowledge or the understanding of SA. They felt “clearer” (F10) about the SA items and became more confident and well-articulated. For example one of the participant F10 said “I feel it much clearer and I can understand it now after discussing with you...if we had discussed at the beginning, I think I would do much better”, which supports the previous study that low prior-knowledge students have the potential to perform similarly to or better than high prior-knowledge students with appropriate help (Yang et al., 2015). Thus, SA is actually teachable and can be improved with appropriate instructions despite the difficulties the students had.

### **Chapter summary**

This chapter has presented the findings in terms of the students’ perceptions on SA, their experience in the SAC assessment, and the challenges they face, which answer most part of RQ 5. These findings informed the understanding of Chinese high school students’ SA

engagement and the focus of improving the students' SAC and revealed the potential positive impact a SAC assessment may have on students' learning.

Specifically, it has been found that the participants had gained understanding about SA by transferring their previous life experiences and the learning experiences in other school subjects. This existing understanding lacked an emphasis on the social aspects of SA especially those that focus on collaborative communication. They generally recognized the value of engaging in SA and had a relatively positive experience of taking the assessment and an expectation of integrating SA into school education. However, they held a conservative opinion on the practical value of SA in improving the possibility of meeting the educational expectations from a current education context (expectations as perceived by them). Moreover, taking the assessment brought them more understanding about SA and the competences that were needed for SA. The difficulties of engaging in SA, that were brought by external context and that they had in themselves (part of them being due to external context), have been uncovered as well. The findings in this chapter will next be related to other chapters to address the overall research aims, which will be discussed in the next chapter.

## **Chapter 9. Discussion**

### **Introduction**

This study aimed to explore the assessment of scientific argumentation competence (SAC) and to understand Chinese high school students' engagement in SA. This chapter will analyze, evaluate, and interpret the key findings presented in the preceding Chapters 5 to 8 to illuminate the answers to the research questions that have guided this research. Section 9.1 will discuss the nature of scientific argumentation, whilst section 9.3 will discuss the assessment design for scientific argumentation. These discussions will help answer the Research Questions 1 to 4 in this study (see section 1.2). In between section 9.2 will discuss the possible considerations in terms of equipping high school students with the competence of engaging in SA, which informs RQ 5.

### **9.1 The nature of scientific argumentation**

In Chapter 3, the conceptual understanding and assessments of SA were reviewed as a basis for developing an SAC assessment. Chapter 6 has validated the SAC assessment and discussed how the assessment results shed light on understanding SAC as a learning progression. Thus, this section will discuss “*the hybrid nature of SA as perceived, experienced, and demonstrated by Chinese high school students*”. This section will first discuss understanding SA from a competence perspective, then will discuss understanding SAC as a learning progression.

#### **9.1.1 Understanding SA from a competence perspective**

As discussed in section 3.2, the limited research in the field of SA that explicitly discusses the competence of engaging in SA perceives argumentative competence as the ways in which different types of skills related to argumentation are manifested in a person's performance (Rapanta et al., 2013, p. 488). Recent research that explores higher order thinking skills and explicitly discusses competence take competence as “the internal structure of competence in terms of basic abilities” (Wang & Song, 2021, p. 694), or “dispositions that are acquired and needed to successfully cope with certain situations or tasks” (Koeppen et al., 2008, p. 62, as cited in Reith & Nehring, 2020). These studies imply that competence is a context-specific construct that requires different dispositions depending on the situations and that competence can be trained and required. In addition, all these studies have approached competence based on the cognitive abilities that constitute it.

Similarly, this study perceived SAC as the abilities students need in order to successfully engage in SA activities. Based on this understanding and the review of the meaning of SA, this study deconstructed SAC into **three components**, namely 1) the ability to identify SA elements, 2) the ability to evaluate SA elements, and 3) the ability to generate SA elements. Based on Toulmin's argument pattern, SA elements include *claim, evidence, reason, and rebuttal*. In this way, instead of assessing the whole argument product that the students generate or assessing students' performance during the whole process, each element of SA competence can be understood and assessed separately. By doing so, this study has aimed to know more in terms of how each competence element contributes to the engagement in SA, and to make SA assessment thereby the instructional guidance based on its assessment more feasible. Chapters 5 and 6 have shown that these competences represent a good part, if not all, of SAC. This finding provides evidence for the conjecture in the previous studies that the competences involved in evaluating argument and systematically identifying argument elements might be part of a common construct (Von der Mühlen et al., 2016; Britt et al., 2014).

Section 3.3 has shown how different frameworks of analysing/assessing SA focus on different aspects of SA (i.e., structure, content, epistemic understanding, and learning progression) therefore bringing the challenge of how to integrate the various aspects of SA into its assessment. Approaching SA from a competence perspective as this study does seems to realize this integration to some extent. The **structure** of SA is included in the assessment because it is entailed in the SAC elements, the **content** of SA is involved by considering the different proficiencies on P-SA elements (as uncovered by the scoring rubrics), the **epistemic** aspect of SA is embodied in the component of E-SA, and all of these together constitute a **learning progression** of SA as shown in section 6.5.1. Therefore, deconstructing SA into competences does appear to facilitate the integration of various perspectives on SA, although the conceptualization of SAC in this study does not capture all of what is meant by SA (e.g., the dynamic social process of engaging in SA).

As discussed in Chapter 3, **content knowledge** is necessary but is not sufficient by itself to support SA engagement. Likewise, Chapter 6 showed that in general, items that need more content knowledge are more difficult for the students, and Chapter 8 showed that participants felt they would feel more confident in engaging in SA if they had more content knowledge. The correlation between content knowledge and SAC is aligned with the notion that competence is context specific (Leutner et al., 2017), indicating SAC should be considered as

specific to a scientific context and to an argumentative context. Nevertheless, Chapter 7 also showed that there was no strong relationship between the students' school Physics test scores and the SAC test scores, although this evidence maybe a bit weak since the content knowledge embodied in the Physics test is different as what is in the SAC test. Students M10 and F12, who showed deep understanding and high engagement in SA in the interview and impressive performance in the SAC test, mentioned that they didn't have outstanding achievement test scores in Physics. This corresponds with the findings of Wang and Buck's (2015) study who examined the relationship between Chinese middle school students' SMK (Subject-matter knowledge) achievement and argumentation engagement. Interestingly, they found that medium-SMK students showed better understanding in terms of viewing SA as a knowledge construction process and had greater potential in argumentation, although sometimes they tended to cite more inaccurate knowledge in their arguments. These findings indicate that content knowledge plays a more limited role in engaging in SA, thus suggesting that SAC is far from being determined by content knowledge proficiency alone.

Acquiring **epistemic understanding** of SA, namely knowing what counts as (good) argument, is significant for engaging in SA and enhanced epistemic understanding of SA leads to the construction of better/more complex arguments (see 3.2.2). The epistemic understanding of SA is entailed in the activities of evaluation and critique (Kuhn et al., 2013; Leung, 2020; González-Howard & McNeill, 2020). This study revealed that it is plausible to include the ability to evaluate SA elements as a component of SAC, while students' knowledge of the meaning of SA cannot predict their *use* of that epistemic understanding. As shown in Chapter 8, some participants could not evaluate the SA element in a specific context even if they knew what the element should be like generally. Similarly, some participants could not generate satisfied arguments even if they could clarify what a good argument should be like. Moreover, Chapter 7 found that students who were provided with the scaffold (i.e., SA elements and their meaning) did not show better performance than those who were not, although almost all participants in Chapter 8 talked about how scaffolding helped them understand SA. It turned out that students didn't tend to apply their understanding about SA deliberately when engaging in SA. Thus, this study revealed the need for a more comprehensive and specific understanding in terms of how should 'the epistemic understanding of SA (Chen et al., 2019)/ the epistemic knowledge of SA (Shi et al., 2021)/ the epistemic work (González-Howard & McNeill, 2020)' be conceptualized. Specifically, the epistemic understanding of SA should not only include knowing the meaning of SA, but also knowing the mechanism of SA. Furthermore, given SA

is an epistemic *practice*, both students' epistemic understanding of SA (i.e., knowing what is (good) SA) and their **application of this understanding** (i.e., applying the knowledge to evaluate/produce arguments intentionally) are important and both need to be treated deliberately. Nevertheless, a question that arises here is whether it is because of an inadequate epistemic understanding of SA, or because it is difficult to apply this understanding to a particular context, that students exhibit the above inconsistencies. A possible explanation could be that deeper understanding about SA and the ability to apply this understanding are reciprocal, while a preliminary understanding (or awareness) is the basis for applying the understanding in SA thereby for a deeper understanding.

Some other findings in this study such as the role 'language', 'perceptions on SA', and 'social skills' play in SA and the discussion of the position of SA in science make it worthy to reconsider the constituents and the scope of SAC. Therefore, it seems necessary to go outside of the literature related to SA to see what 'competence' means and how it can be constructed. Le Deist and Winterton (2005) reviewed various kinds of understandings of competence and pointed out the need for a multi-dimensional framework for competence. Although their review focuses on job qualification, it also informs the conceptualization of competence in education. Their framework includes cognitive competence (i.e., knowledge and understanding), functional competence (i.e., skills required by an occupational area), meta competence (i.e., learning to learn), and social competence (i.e., behaviour and attitude). Similar discussions can be found in Hager and Gonczi (1996) and Weinert (1999). Thus, competence does not necessarily only include cognitive/metacognitive-related abilities.

So, what other aspects could be incorporated into SAC? As shown in Chapter 8, participants indicated that they felt **language** was important to their participation in SA. Likewise, Cikmaz et al. (2021) found similar growth patterns in the argument quality and the quality of language use among fresh college students in a semester's written chemistry lab reports, suggesting that language may be an important lever in support of argumentation. Likewise, Yaman (2020) found that pre-service teachers' argument and use of representations in argument showed parallel patterns of improvement during a two semesters' intervention of argumentation activities in Chemistry. It has been discussed among science education scholars that knowledge cannot be constructed without language and language plays a central role in the development of thought (Norton-Meier, 2008; Wellington & Osborne, 2001; Tang & Moje, 2010). Moreover, argumentation is "a particularly important aspect of the language practice of science"



(Cavagnetto, 2010, p. 337). As discussed in Chapter 3, SA is by nature a social activity, in which people interact implicitly or explicitly with themselves or other people to construct/enhance knowledge. Language is inevitably a medium for the expression and transmission of thoughts. Moreover, language is not only what is said, heard, or written as a product, it is also a process by which ideas are constructed and modified in one's mind or in communications (Cikmaz et al., 2021). So, language is not only interrelated with the argumentation practice, but also part of it.

However, few studies have explored students' use of language within scientific argumentation. Language supports but can also impede the generation of scientific knowledge as discussed by Jiménez-Aleixandre and Erduran (2007). The participants in this current study didn't imply anything related to the role language plays in enhancing understanding and generating knowledge, although they realized that language is needed when building an argument. Another interesting point is that, as showed in section 8.4.2, students like F18 and M4, exhibited performance on the test paper that was much better than what they demonstrated in the interview. Both realized that they found it harder to talk about their ideas than to write them down. In contrast, student F12, who showed excellent performance in both the test and the interview, thought that he could do better when talking about his ideas. Another student, M9, expressed her ideas better in the interview than in the written test. It has been discussed about the various forms of language in science, such as listening, writing, talking, using figures and numerical data, and the different functions they serve in science learning (Rivard & Straw, 2000; Wellington & Osborne, 2001). Therefore, it seems worth noting how the use of different forms of language might support different ways of engaging in SA, and if so, why this happens.

As mentioned earlier, this study focuses on written argumentation, and the **social aspects** of SA were reflected in the assessment in an implicit way by providing dialogues and conflicting ideas for students to compare. As revealed in Chapter 6, Erb items that need to attend to other's argument, and similar Prb items are more difficult for the students. Chapter 8 revealed that the social aspects of SA (e.g., sharing ideas with others, evaluating others' ideas, revising one's own ideas during the interaction etc.) were often absent when the participants talked about their previous understanding of SA, and they realized how they themselves tended to ignore others' ideas. Similarly, the participants thought they would be very nervous if arguing with others face to face. It seems that the participants didn't understand how engaging in SA socially can benefit the construction of an argument and the generation of knowledge, although they started

to be conscious of the social aspects of SA.

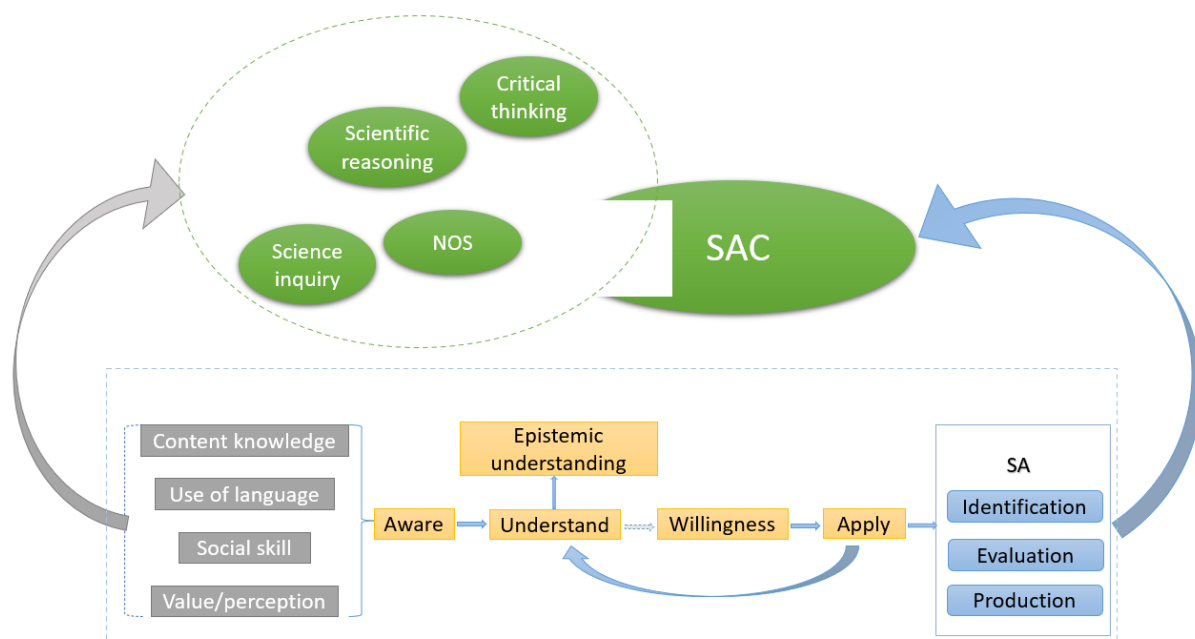
Perhaps precisely because of their lack of understanding of SA, **students' perceptions of the value of SA** also seemed to affect their engagement. As revealed in Chapter 8, students who didn't appreciate the value of SA or who were unwilling to do the test might not choose to submit their test paper. Although most of the interview participants showed willingness in engaging in SA, those who appreciated the value of SA in a way that understands how SA contributes to knowledge construction were more actively involved in engaging in SA. This is in accordance with previous study that found a significant positive relationship between students' being willing and being able to participate in SA (Bathgate et al., 2015). Furthermore, willingness to engage in SA should be the result of an appreciation for the value of SA, regardless of the extent of such appreciation, but reluctance to engage in SA may be due to a lack of understanding of SA or for practical reasons (i.e., limited time, no contribution to their standardized tests performance). Whatever the reason, **willingness** is essential for the effective and sustainable development of SA competence.

The understanding about how the nature of science plays a role in the practice of argumentation as well was interesting. Some of the students' performance uncovered their lack of knowledge in terms of nature of science, such as relying on intuition. The relationship of nature of science and argumentation has been discussed by previous studies, indicating engaging in SA benefits the students' understanding of the nature of science and vice versa (Osborne, 2012; Khishfe, 2020). Until now I have talked about the abilities that were revealed from this study and previous studies as correlated with SA. However, SA has also been discussed with other practices/abilities such as critical thinking (Jiménez-Aleixandre & Puig, 2012), scientific reasoning (Zeidler & Sadler, 2007), and science inquiry (Llewellyn, 2013).

So, to what extent should the meaning or the constitutes of SAC be expanded? A recent discussion by Allchin and Zemplén (2020) proposed the question of where SA should be placed in science education, considering the paramount role it has been given by science education research over the years. They pointed out that argumentation cannot displace all other nature of science education and cannot be taken as a substitute for understanding how science works. For instance, in this study, evidence was provided for the students in test items, and they were not asked to generate evidence by themselves. However, evidence itself does need to be produced and justified. The generation of scientific knowledge requires multiple layers of practices and justifications, and argumentation may happen across the whole process. But

scientific argumentation alone is not enough for actually doing science, and it is not only SA that students need to be taught. Thus, the conceptualization or modelling of SAC should consider sufficient aspects of SA but should also have a scope. The determination of the scope could be based on what is expected for the students in science learning, and the goal that is set for science education.

The above discussion reveals the complex nature of SA, which corresponds to what Jiménez-Aleixandre and Erduran (2007) have argued that different fields (e.g., social culture perspective, language studies, epistemology, philosophy and developmental psychology) can help in promoting SA understanding and vice versa. This section will end with proposing a possible model for understanding SAC based on the discussion. Recent studies in science education, including the current study, usually care about the outcome of or the behaviours shown when possessing a competence (Reith & Nehring, 2020; Romine et al., 2020), which are also referred to as output competences (Cockerill, 1989). However, in education, we should not only care about outputs, but also inputs. Namely, what can educators do to help the students possess or improve their competence. Thus, the understanding of SAC after conducting this study not only considers ‘what student should be able to do to engage in SA’, but also ‘what is required’ in order to be able to do SA. Figure 9.1 shows the understanding about SAC based on the above discussion.



*Figure 9.1 Understanding of SAC revealed by this study*

The above discussion revealed as many possibilities for understanding SAC as the questions it raised. Such as how each aspect interacts with each other and with SAC as a whole, and to what extent should the understanding of SAC be expanded/confined as appropriate for the goal of science education and for its operation. The four aspects in grey are the general competences required to perform SA, but these aspects will also be relevant to other practices. In other words, we are concerned with general competences manifested in the context of SA. To successfully make use of these general competences in the context of SA, students should be aware that these competences are required for SA, then they should understand why and in what ways the competences contribute to SA engagement (this forms their epistemic understanding of what counts as SA and of how SA operates). As discussed above, if the students have a deeper understanding about SA and its value, it may lead to their willingness to engage in SA. But students' willingness to engage is not only decided by their understanding of SA (will be discussed in section 9.2.2). In addition, understanding does not necessarily mean application, so students should be able to apply their understanding in practice (which in turn enhances/modifies their understanding) to identify, evaluate, and produce scientific argument. As shown in Figure 9.1, other scientific practices/competences may relate to SA, but a scope is needed for the construction, investigation, and instruction of SAC. The loop for SAC is not closed because other aspects may need to be added.

Overall, given the hybrid nature of competence, approaching SA from a competence

perspective not only enables the integration of different aspects of SA into its assessment and the investigation of the behavior that should be directly manifested in it, as this study does, but also enables the exploration of the required support behind that behavior. In the context of education, we expect outcomes from students, but more importantly, we should also focus on what kind of inputs can help students achieve these outcomes because this is the function of education. Additionally, approaching SA from a series of competences, expanding while confining the model of SAC has the potential to provide a generic framework that characterizes SA across different science content areas and thus equip students with both a basic and comprehensive understanding of SA (Sampson & Clark, 2006).

### **9.1.2 Understanding SAC as a learning progression**

Assessing scientific argumentation from a competence perspective provides more possibility and feasibility for exploring SAC as a learning progression. Section 6.5.1 has shown a possible learning progression of SAC derived from assessing it. This section will discuss the progression levels of SAC uncovered in this study by comparing with previous studies that have investigated learning progressions of scientific argumentation. In so doing, this study tries to provide some evidence and thinking on making different frameworks of assessing SA comparable.

The learning progression shown in Chapter 6 is generally consistent with Osborne et al.'s (2016) learning progression for argumentation. Osborne et al. (2016) proposed a learning progression with three general levels (i.e., level 0, 1, 2) and corresponding sub-levels (i.e., sub-level a, b, c, d). The comparison between the two progressions is shown in Table 9.1. Although 'Construct claim' and 'Identify claim' were not included in the final version of the assessment in this study, as the test items of these two indicators showed extremely low difficulty in the pilot studies. Despite the general consistency, this study further partitioned indicators of argument construction (i.e., Provide evidence-0c, Construct reason-1a, Construct an argument-1c) based on the relevance, accuracy and coherence of the argument elements as embodied in the scoring rubrics, which revealed the non-linear relationship between I-SA, E-SA and P-SA. Moreover, Osborne et al. (2016) assigned 'Identify evidence' to level 0c and 'Identify reason' to level 1a, but this study did not find differentiation between these two indicators.

*Table 9.1 Comparison of learning progressions*

<b>Competence level in Osborne et al. (2016)</b>	0a Construct claim	0b Identify claim	0c Provide evidence	0d Identify evidence	1a Construct reason	1b Identify reason	1c Construct an argument	1d Provide an alternative counter argument	2a Provide a counter-critique
<b>Competence level in This study</b>	1	1	1-2	2	1-3	2	1-3	2	3

This study not only further verified the learning progression proposed by Osborne et al. (2016) but expanded it by adding the *Evaluation of a scientific argument* (E-SA) into the progression. There is a pseudo-contradiction between the two learning progressions: Osborne et al. (2016) took ‘critique’ as the highest level of their learning progression by including sub-level 2a, 2b (One-sided comparative argument), 2c (Two-sided comparative argument), and 2d (Provide a counter claim with justification); while all the E-SA elements in the learning progression in this study were located at level 1 and level 2. Although Osborne et al.’s (2016) critique may include implicit evaluation of SA, and students’ evaluation of SA in this study may include implicit critique, they are actually different. The E-SA in this study resonates to the *checking* sublevel within the *evaluate* level in Bloom’s Taxonomy (Anderson & Krathwohl, 2001), students need to evaluate each SA element based on the provided criteria by implicitly judging the connections between certain elements, however, they don’t need to explicate and explain their judgement or to evaluate competing arguments on a controversial topic. However, the items for Osborne et al.’s (2016) critique level all require either implicit or explicit evaluative judgement to form an argument. Following the focus on argument evaluation in the current study, if we extract the evaluative part of Osborne et al.’s (2016) critique and to make it explicit and deliberate, such as asking students to provide warrants for their judgement or generate evaluation for one or more arguments strategically based on the norm of argument, it may be more difficult than the E-SA in the current study given students need to explicate the connections and make deliberate evaluation (Britt et al., 2014). Therefore, a more comprehensive learning progression of SAC may be generated if combining learning progressions in the current study and in Osborne et al. (2016) and including asking students to generate evaluations of SA.

As mentioned in section 5.4.2, a social scientific issue task (Task 6) is included in the test, and the test tasks within each component have different complexities according to the required content knowledge, involved information, and students’ familiarity of the topic. Each P-SA element in Task 6 was found easier than items in the other scientific-based tasks. Task 5 includes more information, and the topic is unfamiliar for students. It turned out that each P-SA element in Task 5 is more difficult than in Task 4 even though Task 4 requires more content knowledge. Task 7 includes more information and more content knowledge (i.e., need to use basic formulas) but it is a familiar topic adapted from the textbook, P-SA elements in Task 7 turned out to have similar difficulty with those in Task 5. Although our assumed difficulty of

the three components is confirmed in general (i.e., I-SA is the easiest, E-SA is easier than producing coherent and plausible arguments but more difficult than producing simple and relevant arguments), when students need to coordinate multiple pieces of information or use formulas to generate even simple connections between claim and evidence, this seems more difficult for them than evaluating arguments (Pr\_5.2 and Pr\_7.6). Moreover, interview participants in Chapter 8 mentioned how they felt it demanding to deal with too much information. In general, the nuanced differences in the difficulty of each P-SA element between tasks is generally consistent with our assumed complexity of each task (i.e., Task 6 < Task 4 < Task 5/Task 7). Although the current study considered the content knowledge, information and familiarity when deciding the complexity of each task, this study didn't control for each of them separately and test their effects directly as this would require a far longer assessment. The size and appropriateness of data that should be used in items was discussed in Berland and McNeill (2010) learning progression, similarly, Osborne et al. (2016) also reported that the relative number of claims and pieces of evidence increases task difficulty. Additionally, previous studies discussed the necessity of treating content and SA separately (Yao et al., 2015). Future studies can therefore explore learning progressions including content, size of information, and topic familiarity explicitly and separately to reduce the variance when assessing SA.

The SAC progression in this study also suggests that attention when exploring students' SAC should be paid to the extent to which students can apply their epistemic understanding of SA, in addition to whether they have epistemic understanding on SA or not. Findings in section 6.5.1 showed that even for the same E-SA element, it is easier for students to evaluate when the context is less complicated. This resonates with the discussion in the previous section that students need to be *aware* of and *understand* SA as an epistemic practice and to *apply* this understanding. The influence of the nature of each SA element on the complexity of its evaluation is reasonably consistent with previous studies. Compared with evaluating whether the evidence is relevant (Ee\_2.1; Ee\_3.1), it seems more demanding for students to evaluate whether it is sufficient (Ee\_7.3). This corresponds to the findings that students tend to provide a single piece of evidence rather than considering evidence comprehensively (Shi, 2020). In parallel with the awareness of the importance of understanding others (Romine et al., 2020), it is easier for students to recognize that a rebuttal does not attend to other's argument (Erb\_3.4) than to evaluate whether the rebuttal weakens other's argument (Erb\_2.2; Erb\_7.2). This study thus responded to Sampson and Clark's (2006) call for exposing the standards used to construct



and evaluate arguments by the scientific community to students by including the standards in the learning progression and analysing students' understanding of each standard explicitly.

Some inconsistencies were found in the P-SA items: the scoring rubrics for Prb items were 0-3 under the SSI context while they were 0-2 for items under the scientific context. This was because the original scoring categories of 'does not pay attention to other's argument' (i.e., score of 1) and 'attending to other's argument but not accurate nor coherent' (i.e., score of 2) were not differentiated in Prb items under scientific context. As mentioned in section 6.4.2, the students' engagement in rebuttal seems to be highly dependent on whether they thought it necessary to attend to other's argument and how easy it was to do so. Students viewed Task 6 as an open question thus they seemed more likely to ignore another's argument in cases where there were no wrong statements in the opposing argument. Since only three items were designed to ask students to engage in rebuttal, it needs further exploration in terms of how students' engaging in rebuttal may be different in different situations.

Overall, the alignment of the learning progressions between this study and previous studies suggests that understanding SAC as a learning progression is plausible and has the potential to make different studies comparable. Combining with the discussion of understanding SA from a competence perspective, it seems worth considering how to incorporate other aspects of SAC into one or may be several learning progressions. Furthermore, this section also revealed how the task context (i.e., provided information, topic familiarity, required content knowledge) influenced the difficulty of items. Thus, the consideration of SAC may need to consider the various component competences that constitute it, and how each component competence correlates with each other to form a learning progression, and how the context in which SA happens may support or impede students' engagement in it.

## **9.2 Equipping students with SAC by assessing it**

The previous section discussed the plausibility of taking SA to be composed of various competences and exploring SAC as learning progression(s). This section will discuss the characteristics of Chinese high school students' understanding and performance of SA together with the current educational context that has shaped it. By doing so, this section argues that *"Chinese high school students' SAC needs to be improved, and assessing SAC has the potential to add to its teaching and learning"*.

### 9.2.1 Chinese high school students' SAC

This subsection draws together findings from Chapter 6 to 8 to discuss how SA was perceived, experienced, and demonstrated by Chinese high school students. Previous studies have shown the difficulties students have in performing skillful argumentation, and that students have an ability to argue from an early age when provided with appropriate prompts (Bricker & Bell, 2012; Kuhn & Udell, 2003; Stein & Miller, 1993). Likewise, the cohort of Chinese high school students participating in the interviews appeared to have a general understanding of SA that was transferred from other subjects and life experiences that explicitly talked about/implicitly involved the practice of argumentation, despite there seeming to be a lack of explicit SA instructions and practices in their science classrooms. Consistent with most of the students' rather shallow understanding of SA, section 6.5.2 revealed that the majority of students who participated in this study were at lower levels of the SAC learning progression. So, it seems of particular importance to engage students in argumentation in the context of school science education deliberately and explicitly. The rural-urban disparity in education in China has been mentioned in section 2.3. Similarly, PISA 2018 found that there was still a big gap on students' performance between urban and rural areas after controlling for their social-economic background even in the more developed provinces of China (OECD, 2020). Thus, although the sample in this study does not represent the overall situation in China, the average performance of high school students in the whole country seems unlikely to be better than the results of this study given no schools from rural areas participated in this study (see section 4.3.1).

A surprising result was the positive attitude to SA that students expressed in the interviews, which resonates with studies conducted in other countries that student held positive attitude for SA (see section 3.4.3). Chapter 8 showed that almost all the participants held a reasonably positive attitude towards SA. Although most participants' positive attitude was based on the SAC assessment, some of them appreciated the value of SA for their science learning and thinking in a deeper way. However, they became pessimistic when talking about the integration of SA in science classrooms. The participants at their age (16/17 years old) already had a relatively comprehensive understanding about the broad social and educational context they were in. Therefore, their attitude toward SA seemed to be twofold: they enjoyed engaging in SA and even appreciated the value of SA, but they thought it is not 'useful' for the aim of high school education as perceived by teachers, schools, and society. These young students seemed to have developed an instrumental approach to learning, which was focused on examination

success, and they tended to take what they think the outside world expects of them as their own expectations of themselves. Previous studies have talked about the importance of letting students appreciate a scientific practice to better engage in it (Chen et al., 2019; Bricker & Bell, 2012). But for Chinese high school students who have become cognitively mature to realize the value of SA, it seems both the intrinsic value and the practical value of SA matter for them. This resonates with the discussion in section 2.3 that Chinese students showed contradictions in viewing the aim of education with its intrinsic value while viewing their current school education with extrinsic values.

Even if they did not possess strong content knowledge and epistemic understanding of SA, the students could still engage in SA. Chapter 6 showed that it was harder for the students to evaluate and generate SA when more content knowledge was required or under complex context. However, they can engage in the practices of evaluation and production of SA when less and simpler content knowledge is demanded or when the context is less complex. This indicates that students can still participate in SA even without strong content knowledge. Similarly, parallel with the recent call for letting students know ‘what counts as’ SA and ‘what counts as’ good SA (Chen et al., 2019; Groom et al., 2018; Sampson & Clark, 2006), Chapter 8 revealed that even if students do not have a thorough understanding about what SA is, they can still have some sense of what good SA should be. As shown in the learning progression reported in section 6.5.1 that the students can evaluate simple arguments (Level 1) even if they cannot differentiate between evidence and reason (Level 2) and cannot generate a plausible argument (Level 2). Participants in the interview also talked about their understanding of what they thought good SA should be in a quite reasonable way. Therefore, students can participate in evaluation and critique even if they have not understood what SA is thoroughly.

A prominent characteristic in the students’ SA is that they need to build more ability in the social aspects compared to cognitive aspects of SA, and similarly in evaluating knowledge compared to acquiring knowledge. As shown in Chapter 8, the participants tended to have a much weaker understanding about the social aspects of SA, which was revealed in their performance as well. For instance, some students’ understanding of arguing with other people was prone to be “socially undesirable quarrelling, partisan bickering, or intractable articulation of differences in opinion” (Bricker & Bell, 2012, p. 127). Additionally, the students tended to ignore the existence of multiple voices within the same problem and pursue the only ‘correct answer’. One could argue that students’ tendency to ignore other voices may be due to the lack

of test items with multiple possible answers in the SAC assessment. However, as shown in section 6.4.2 and discussed in section 9.1, the students seemed to be more likely to attend others' argument in their rebuttal in items under a science context (which has one correct answer and there are mistakes in the provided arguments) than the item under the social science issue context (which does not have a correct answer and only provides opposing claims). This implies that the students accept that people can have different ideas on the same problem but then ignore evaluating/comparing different solutions when the problem is open and there is nothing 'wrong' in other's argument. As mentioned in section 3.2.2, Kuhn (2000) categorized epistemological development into three stages: absolutists view knowledge as fixed, certain, and independent from human cognition, multiplists take knowledge as subjective and depending on personal experience without a need for reason and compare, while evaluativists consider knowledge as a result of examination, comparison, and evaluation. The students' understanding and performance on SA showed that they tended to consider all ideas to be equally valid (Sengul et al., 2020), which is consistent with the multiplists category. However, to successfully engage in SA and science learning, we expect students to be evaluativists.

Another characteristic is the students' limited ability to compare and interpret data. A prominent pattern shown in the interview data is that the students tended to either ignore the provided information or accept all the information without evaluating it. It seemed that some of them didn't view information as resources to be used to help them solve problems. This again revealed their limited ability of evaluation, further revealing their deficient epistemic understanding of SA, specifically of the value of SA.

Overall, the group of Chinese high school students' attitude about SA was twofold: appreciated its intrinsic value but was pessimistic about its practical value. They demonstrated existing understanding and ability of engaging in SA, indicating they have the potential to actively engage in SA when provided with appropriate environment. However, students' awareness, understanding, and practice of SA were still limited, especially their ability to evaluate knowledge and generate knowledge socially. The current curricula, as discussed in section 2.1, provides very simple and unsystematic descriptions of SA in the achievement/grade progressions. Despite that, to help students achieve the scientific thinking desired by the curricula, classroom interventions are needed to develop students' SAC.

## 9.2.2 What to expect from assessing SAC?

The above discussion suggested that the group of Chinese high school students' SAC needs to, and has the potential to, be improved and SA implementation needs to be facilitated in school science classrooms. This section will discuss how SA assessments may influence SA teaching and learning from the impact and content of assessments. I'll start by discussing why changing assessments, especially high-stakes examinations, is critical to implementing SA practice in science classrooms. Then I will discuss how assessing SA may influence students' science learning in a positive way.

Section 3.1 has discussed that high-stakes examinations usually lead to exam-oriented teaching and test results-focused learning. Similarly, the participants in this study expressed how the focus of achieving high examination scores had shaped their learning experience and their pessimistic view of the integration of SA in their classrooms (see section 8.4.1). The findings in the current study resonates with the discussion in Chapter 2 in terms of the exam-oriented culture and the Gaokao-focused education in China. Given high school students in China are under the pressure of entering into colleges, the college-entrance examination (Gaokao) is their primary focus and priority. The increased focus on achieving high scores on the Gaokao has made teachers and students short of time to do other activities that explore the process of science learning. Under the pressure of Gaokao, learning, especially in high school, seems to have become exclusively repeating as many exam questions as possible (Yu et al., 2018; Liu & Helwig, 2020). As has been recognized by many researchers in China and in other countries, what is emphasized in classrooms is what is measured by examination items (Osborne et al., 2016). It seems especially the case for high schools in China that teachers only teach, and students only learn what may appear on the Gaokao papers. The previous section has discussed that both the intrinsic and practical value of SA matters to Chinese high school students. Therefore, in order to integrate SA into the science classroom and for students to appreciate its value, it needs to be assessed.

Similar to the students' focus on the practical value of SA, section 2.3 has discussed the fact that Chinese students tend to relate high school education with extrinsic values such as doing well on the Gaokao, entering into college, and finding a well-paying job (Liu & Helwig, 2020). However, as argued by Yu et al. (2018), school education should have "both intrinsic and extrinsic merits" (p. 204). Education should show students, through teaching and assessment, what is appreciated by the intellectual community and society at large and what is important to

students' life-long development. As revealed from the students' perspectives in Chapter 8, the current examinations have not been aimed at SA and seem even to impede the development of abilities appreciated by SA. The impacts include making the students form fixed mindsets, being afraid of asking why, focusing on results and the single correct answer. Moreover, the students themselves had a clear awareness of all the disadvantages of the current examinations and the undesired learning experiences shaped by them. Most importantly, most of the students seemed to have their thoughts on what they need and want to experience in school science learning, such as to learn what is relevant to their life and to be engaged with critical and evaluative tasks such as argumentation. Despite this, they still have to conform to what is imposed upon them. Thus, it seems that the current examinations have shaped the students' science learning experience in a way that meets neither students' own needs for learning nor what is appreciated by the goal of science learning in the curricula. The school education in China, therefore, seems to have made students well aware of the extrinsic merits of receiving education, whilst it does not appear to provide many opportunities for students to experience its intrinsic value.

However, as revealed in Chapter 8, students appeared to have relatively positive experiences with taking the SAC assessment. Although the assessment itself is really a first step, but here rather thinking through the lens of their experience of taking the SAC assessment, we can uncover what students appreciate and the potential of such assessments to positively impact their learning experience. As mentioned in Chapter 8, most students found the assessment engaging because it is close to real life, which is consistent with the assertion in previous studies that students are more motivated when they find their work meaningful and relevant to their life (Berland et al., 2016). It also resonates with an appreciation found in the literature for the recent Gaokao to include more real-life scenarios and scenarios related to real science research (see section 2.3). Moreover, students reported that the lack of connection to real world problems of the current test items and lack of practices in science learning made them feel discouraged from learning, although they can still get a fulfillment from achieving higher scores than others. However, this kind of fulfillment is far from what is expected from learning.

In addition, students' positive experiences came from feeling like they've learned something new. Repetitive ways of learning and assessment may cause students' inflexible expectations toward learning and assessment, which can lead to inflexible skills development (Crisp et al., 2008; Koretz et al., 2001). However, assessments that do not always conform to students'

inflexible expectation can bring novelty, unfamiliarity, perplexity, ambiguity, or complexity, which brings uncertainty for learning thus engendering learning (Chen & Qiao, 2020). Thus, assessments should not be all repetitive knowledge/forms, especially in cases where practicing exam items has been a major way of learning.

This inflexible expectation and skills development were also revealed from this study in that the students found the assessment novel but difficult compared with the normal tests. However, as discussed in Chapter 8, the students enjoyed the experience of really thinking about a problem rather than thinking about how to use formulas to obtain the right answer. When the students care less about the final answer and pay more attention to the process of learning, they would also be less afraid of getting the wrong answer or saying something wrong, as mentioned in Chapter 8, thus they would be more actively engaging in knowledge construction. Therefore, assessments that focus on various scientific competences and show the “often tacit epistemological commitments” (Sandoval, 2003, p. 8) established by the science community can provide students with the opportunity to develop flexible skills that are appreciated by the community and meet students’ own expectations of learning.

Now we return to the argument of assessing SA in examinations. A positive relationship between students’ being willing and being able to participate in SA has been found and students who held a negative value towards SA gained significantly fewer benefits from engaging in SA (Bathgate et al., 2015). Therefore, students’ willingness to participate in SA is important. As discussed previously, Chinese students’ perceptions about education seem to have been shaped as focusing more on its extrinsic value, although they know its intrinsic value. However, students in the current study doubted the practical value of SA to help them obtaining higher examination scores although they were in general aware of the intrinsic value of SA. So, assessing SA in examinations can eliminate students’ doubt for the extrinsic value of SA thus to improve their willingness of engagement. Vansteenkiste et al. (2009) found that framing learning activities based on intrinsic goals (e.g., self-development and community contribution) can bring more benefit to the students compared to extrinsic goals in terms of the performance, engagement, and experience in learning. Thus, by assessing SA, SA seems promising as an activity that students perceive as having both intrinsic and extrinsic value.

As mentioned in section 2.3, although SA is not included, the current Gaokao items generally have high demands on students’ understanding of content knowledge and reasoning skills. However, previous studies and the current study showed that students tend to view it as

focusing on memorisation. So, students' perception on the current examination may be due to their use of memorization as a strategy to pass the high-stakes examinations. Combining with the finding shown in section 8.4.1 that a few students used game-playing strategies to answer the multiple-choice questions rather than thinking about it, it seems reasonable to consider whether students would also use strategies such as memorization to deal with examinations (especially when it is high stakes) that include SA. However, as mentioned in section 3.1.2 that assessment can have positive impact on learning and teaching and can convey signals in terms of what matters in education and life, assessing SA would make teachers and students put more emphasis on implementing such activities in classrooms. Therefore, assessing SA may help the curriculum reform be effective and successful (Yu et al., 2018), rather than being "old wine in new bottles" (Zhao et al., 2015, p. 6). Yet, assessing SA can only be part of the effort of integrating SA into school education. Future research still needs to explore how to implement its teaching in classrooms and how assessing SA impact its teaching in practice.

Overall, this section has discussed the students' experience of taking the SAC assessment and school learning. Together with the discussion in section 3.4.3, Chinese students indeed seems to be able to and benefit from engaging in SA, and they also appreciate the value of SA despite their willingness is also based on the practical value of SA. This section argued the necessity of assessing SA in order to teach it and learn it in Chinese high schools' science classrooms and the potential positive impact of SA assessment on learning. The next section will discuss specifically about the development of a SA assessment.

### **9.3 Developing assessments for scientific argumentation**

The previous sections have discussed the complex nature of SA and the necessity to assess SA. This section will discuss the multiple considerations required for designing SA assessments in terms of reducing the influence of construct-irrelevant factors on participants' performance. Specifically, this section will discuss how 'test-takers' perceptions of what is assessing', 'scenarios and its arrangement', 'item format', 'language use', 'the provided information' may influence the validity of a SA assessment. These considerations were mainly generated by conducting the iterative process of designing and validating the SAC assessment, as shown in Chapters 5 and 6. So, a fundamental claim for this study is that "*An iterative process of developing and validating a SA assessment does help generate assessment design guidelines*". This study also argues that "*Assessment design guidelines are needed to facilitate SA assessment*".



Previous studies, especially large-scale studies aimed at assessing SA, have realized that their test takers had never been taught SA explicitly and thus were unfamiliar with engaging in SA practices (Osborne et al., 2016; Deng & Wang, 2017). Participants in the current study also mentioned that they had never been invited to practice SA and never heard about SA in their science classrooms. In this case, the test-takers' perceptions of what is being assessed becomes worth considering. As first mentioned in section 3.1, sometimes what the test designer intends to assess does not match what the test-takers think they are being assessed on (Cheng et al., 2011; Qi, 2007). An assessment that deviates from the usual types of tests can easily cause this mismatch since test-takers' previous experience of taking tests to some extent shapes their interpretation of an assessment (Crisp et al., 2008). This mismatch, from a perspective of the assessment impact, could bring unintended impacts to learning and teaching, whereas for the assessment per se, it could undermine the assessment validity. As argued by Koretz et al. (2001), some amount of instruction for the test-takers is needed to increase familiarity and therefore to improve the validity of test scores. The findings in Chapters 7 and 8 have shown that providing the scaffold (meaning of SA) to the participants did not improve their SAC test scores, but the scaffold indeed helped the participants to gain more understanding about SA and therefore about what the assessment expected of them. Beyond explicit instructions, more considerations are needed to focus the test-takers' attention on SA and to minimize the construct-irrelevant influences.

To engage students in argumentation, as in this study, existing research in science education often employs a scenario-based approach to assess SA (Osborne et al., 2016, Lee et al., 2014). As has been discussed in section 3.3.5, Deane et al. (2019) found that lead-in tasks and the task sequence supported the students' argument writing. Similarly, this study found that by focusing primarily on one SAC component in each task and arranging the scenarios in sequence (i.e., I-SA tasks followed by E-SA and P-SA tasks; simple task followed by complex task within each category that assesses one SAC component, see section 5.4.2), this was helpful to resolve the participants' confusion as they proceeded with the assessment. As presented in Chapter 8, the participants found the focus of each category of tasks clear and expressed that they learned more about the SA process by engaging in the assessment. Thus, it is worth considering scenario arrangement to reduce the test-takers' extra cognitive load and provide instructional information.

Likewise, unfamiliarity with SA may lead us to ask, "Do students know what they need to

argue about when they take the assessment?”. The ‘Make the problem to be argued explicit’ factor identified in section 5.4 had helped students focus on the issue that they need to argue, thereby reducing the threat of unfamiliarity to validity. However, a potential problem often present in scenario-based assessments is item dependence, which is caused by different items using a common scenario (Jiao et al., 2012; Wang et al., 2005). As shown in section 5.4, this needs particular attention in SAC assessment since different SAC elements (i.e., sub-skills needed for SA) are correlated with each other in nature and by the common context.

Item format should be a consideration in SA assessment, but not a constraint. It has been discussed how item formats that allow for extended writing or articulation of reasoning can fit into the assessment of higher order thinking skills (Haladyna & Rodriguez, 2013). Whereas others have criticized constructed-response assessments for not detecting the process and components that contribute to the response and the limited information a single score of it provides (Levy, 2013; Hillocks, 2002; Deane et al., 2019). However, as shown in this study, focusing on the competences that are needed for an activity/process to some extent reduces the limitation of choosing either one item format or the other. In other words, the decision on the item format should be consistent with the competence it is designed to assess. Although multiple-choice (MC) questions have been criticized for their inability to assess advanced literacy skills, Osborne et al. (2016) advocated to include MC items in large-scale SA assessments. However, their study ultimately did not include MC items because of the poor item characteristics of MC items. Therefore, both theory and practice need to be considered when deciding on the item format.

When the assessed construct changes, usually the form of assessment changes as well especially in cases in which the construct changes from content knowledge to higher order thinking skills like SA. As Ng Yee Ping (2019) points out assessments in SA usually set items in real-world scenarios to motivate test-takers. However, the unfamiliarity in the assessed construct and the form of assessment can also threaten test-takers’ psychological safety (Nasir et al., 2006). As shown by the findings in Chapter 8, the participants in this study found the assessment engaging because the construct and item form were novel and the scenarios were close to life, but they also found it difficult for the same reason. The ‘familiarity’ at this point needs to be interpreted according to the broader context, namely, in a school science learning context or in an everyday life context. In other words, scenarios that are familiar in everyday life are not necessarily familiar in school science learning. Ahmed and Pollitt (2001) have

talked about how real-world scenarios lead to increased cognitive workload for test-takers and pose a threat to validity. Therefore, more concrete and targeted considerations are needed for designing SA assessments that are dominated by real-world settings and that students are often unfamiliar with.

Chapter 5 has presented the factors that are worth considering when developing an SA assessment. Some of them are general ones that have been reported by other studies as well. Ahmed and Pollitt (2001) argued that language in a real-world context is usually more complicated than that in context-free scientific scenarios thus reading ability is often needed for understanding the assessment questions. Crisp et al. (2008) have also reported how changing the wording of items reduces the threat to test validity. Findings in Chapter 5 indicated that wording should be succinct and focussed, while in the meantime, the students' previous experience of learning and testing seemed to have affected how they interpret item questions/sentences. Thus, students who have different cultural and education background and experience may interpret assessment items differently especially given SA is essentially a discourse and moderated by language (Erduran et al., 2015). The finding in Chapter 7 that the students' Chinese school test scores correlated positively ( $r = .22, p < .001$ ) with their SAC scores also indicates the importance of language in SA assessment construction. This in turn justifies another SA specific factor found in Chapter 5 that SA-related terms need to be clarified in a straightforward and close-to-context manner, especially if the test-takers have not previously been taught about SA.

The use of irrelevant information/data in a test item has been discussed by previous studies. Berland and McNeill (2010) took instructional context as part of a learning progression in which items including both appropriate and inappropriate data are more complex than those including only appropriate data. In contrast, Ahmed and Pollitt (2001) discussed how the irrelevant information in real-world contexts distract the test-takers' attention thus reducing the validity of assessment. This study found that some participants tended to believe that all the information provided was correct, and some of them found it difficult to sort out irrelevant information when they were provided a lot of information. This corresponds to Crisp et al. (2008) who found that the students expected every piece of information in a test item to be useful, even when it wasn't. For SA assessment, providing irrelevant information in an assessment seems desirable given the students are expected to be able to identify claim-relevant evidence among other information. Thus, what should be considered becomes how much

*irrelevant* information and what kind of *irrelevant* information is needed. This depends on the purpose of an assessment and the group of test-takers it is aimed at. For instance, the irrelevant information could be the students' common misconceptions or something that can lead to new knowledge when an assessment aims to promote students' content knowledge understanding about a topic.

Another related consideration is the content knowledge provided/required by an item. As talked about in Chapter 5, when too little appropriate information was provided and more content knowledge was needed, it was difficult for some students to engage in SA because they didn't know the knowledge. Section 3.4.1 has talked about how previous studies have found a proficient understanding of content knowledge helps students perform better in SA (Yang et al., 2015, etc.). So, it is necessary to control the requirement for the proficiency of students' content knowledge to provide them with the opportunity and possibility for arguing, using and comparing evidence rather than recalling knowledge. Similarly, the point here is how much content knowledge is appropriate for the SAC assessment. To answer this question, it is essential to look at how the targeted participants respond to the items. Based on the aim of an assessment and its target test-takers, provision of the *relevant* information in an item should consider the possibility to allow test-takers to do argumentation.

For now, I've discussed the considerations needed to focus on target competences and reduce construct-irrelevant variance. The caveat here is that if we try too hard to reduce unfamiliarity there may be undesirable results, because all the information provided in the assessment to support test-takers may also impose some limitations upon them. Rockstuhl and Lievens (2021) discussed the pros and cons of using general and specific prompts in a scenario-based assessment and found that an item with more specific prompts is more predictive for cognitive constructs, whereas an item with more general prompts is better at eliciting personality constructs since there are less constraints for the test-takers to express various ideas. Similarly, Crisp et al. (2008) gave a warning to not over-coach because it may cause students' inflexible expectations therefore developing inflexible skills, although they argued to provide basic assessment-related information to the test-takers to avoid contradicting their expectations of assessments (Koretz et al., 2001). So, the assessment design in this study, such as providing scaffold and furnishing criteria for E-SA items, may to some extents also have limited the potential of the assessment to detect extra information in terms of the students' SAC.

By comparing to the existing literature, the preceding discussion highlights the complex nature

of designing an assessment for SA. This study has indicated that the iterative assessment development procedures and argument-based validation promoted the critical understanding of a SA assessment and thus its improvement. It has been a common norm for test design to be iterative and revisions made in response to data from tryouts (AERA, 2018), but the emphasis of assessment studies has been focusing more on the instrument that resulted from an iterative process or the outcomes resulting from using the instrument, making it difficult to use/adapt instruments used in different studies. This further impedes studies dedicated to assessment development from being a sustainable enterprise. This gap can be revealed by the lack of guidelines about the assessment of SA despite the various ways proposed to assess it as discussed in section 3.3 (Ng Yee Ping, 2019).

Therefore, the discussion in this section ends with proposing a set of SA assessment guidelines. As mentioned in section 3.3.5, Ng Yee Ping (2019) proposed a Three-cornerstones model for designing SA items, that is, ‘argumentation’, ‘item anatomy’, and ‘learning objectives’, but she did not systematically deconstruct each cornerstone. On the basis of synthesizing the previous studies, some extra considerations are added to form an expanded assessment guideline for SA assessment. Consistency between what was found in this study and what was discussed in previous studies also reveals the usefulness of adopting an iterative process and probing into test-takers’ experience for developing SA assessments.

As shown in Figure 9.2, four aspects are considered in constructing a SA assessment, each aspect contains several categories that are further specified into several factors. The factor in blue was that explicitly talked about by previous SA assessment studies but not considered in this study, the factors in black were those considered by both this study and previous studies, and the factors in green were those proposed to be used in SA assessment for the first time by this study.

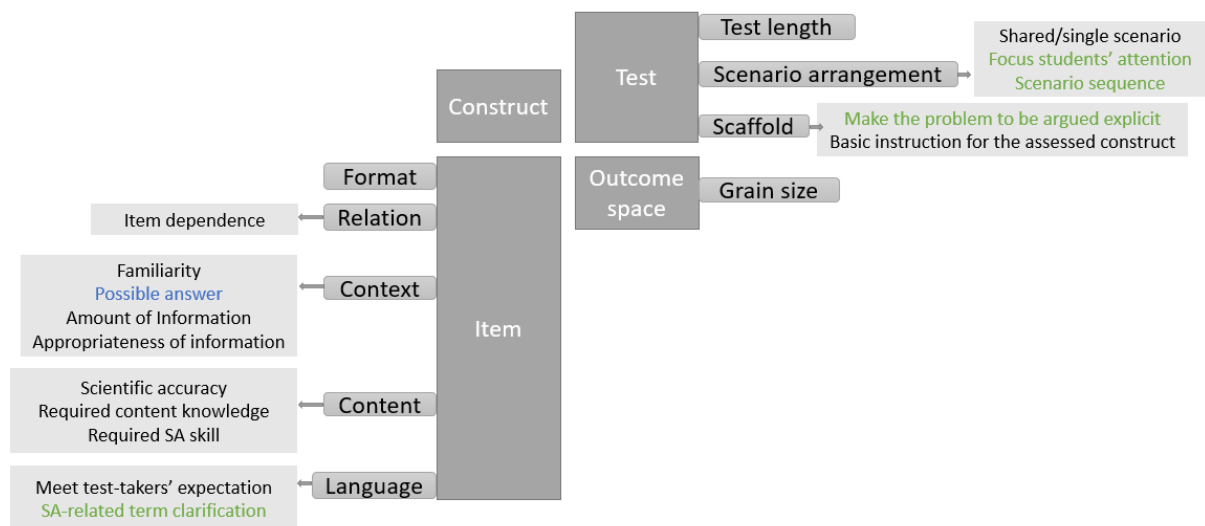


Figure 9.2 SA assessment guideline

These aspects, categories and factors are not totally separate, but are related and may influence each other. For instance, different interpretations of the SA construct may lead to different decisions about these factors. The content assessed by an item may affect how the context is set, and vice versa. Additionally, the process and the product of developing the SAC assessment and the influencing factors found to affect the performance of the assessment may not be universally applied to all the situations of SAC assessment. Depending on the aim and design of a SA assessment, some of these factors may not need to be considered, or maybe more factors need to be included, but the documentation of developing a SAC assessment provides opportunity for those who are interested in the assessment of SA to scrutinize thus provides empirical information to be used in further research.

In the end, this is just a set of guidelines, not a recipe. As highlighted above, developing a SA assessment is complex, and even what is listed in the guidelines should be carefully considered when developing SA assessments in practice as to whether it is appropriate and to what extent it should be used.

### Chapter summary

This chapter discussed four arguments, leading to a more comprehensive understanding toward SA and its assessment as understood, experienced, and performed by Chinese high school students. This chapter first discussed “*the hybrid nature of SA as perceived, experienced, and demonstrated by Chinese high school students*”. Specifically, this chapter discussed the

possibility and advantages of considering SA as composed of a series of competences. The behaviors of engaging in SA, namely the Identification, Evaluation, and Production of SA, have been justified in this study as essential components of SAC. In addition to this, competences that support these behaviors, namely, ‘content knowledge’, ‘use of language’, ‘social skills’, and ‘perception of SA’ matter as well for understanding SAC from an educational perspective. Moreover, what propels students’ successful engagement in SA is determined by their awareness and understanding of how these competences contribute to SA engagement, and their willingness and ability to apply these competences and understandings to SA engagement. In addition, exploring SAC learning progression(s) is plausible given the consistency between the learning progression generated in this study and those from previous studies.

Secondly, “*Chinese high school students’ SAC needs to be improved, and assessing SAC has the potential to add to its teaching and learning*”. Chinese high school students’ SAC performance was relatively low, but they tended to have existing understandings about SA, hold positive attitude towards SA and show the capability of engaging in SA. Both the intrinsic and practical value of SA matter for their engagement in it. They were particularly weak in social aspects compared to cognitive aspects of SA, and similarly in evaluating knowledge compared to acquiring knowledge. The current Gaokao-focused school education seems to have shaped the students’ learning experience in an inevitable while undesirable way. However, the SAC assessment provided a more positive experience for students and has the potential to lead to a learning experience that students desire and is appreciated by the scientific community compared to current exams. Assessing SA may increase students’ willingness to participate in SA by revealing its extrinsic value.

Thirdly, “*An iterative process of developing and validating a SA assessment does help generate assessment design guidelines*”. The findings generated from the iterative assessment development process, which helped improve the SAC assessment in this study, turned out to be mostly aligned with previous studies. Lastly, “*Assessment design guidelines are needed to facilitate SA assessment*”. The considerations of ‘test-takers’ perception of what is assessing’, ‘scenarios and their arrangement’, ‘item format’, ‘language use’, ‘the provided information’ in SA assessment improved the assessment in this study and helped understand the complex nature of designing a SA assessment. Given the lack of guidelines for SA assessment and the need for conducting SA assessment, a set of guidelines that may help advance the area of SA

assessment was proposed.



## **Chapter 10. Conclusion**

### **Introduction**

This chapter will conclude the thesis by summarizing the findings, providing implications, clarifying contributions, and reflecting on the limitations of this research. Section 10.1 will provide answers to the research questions by summarizing the research findings. Section 10.2 will present how the findings of this research inform educational policy and science teaching aimed at Chinese high school students noting that the implications may inform other contexts as well. The contributions of this research will be elaborated in section 10.3. Limitations of the research and possible directions for future research will be discussed in section 10.4.

### **10.1 Answering research questions**

#### **10.1.1 RQ1. How can a SAC assessment be designed for high school Physics students in China?**

An iterative process, as demonstrated in Chapter 5, has been adopted to develop a SAC assessment in this study, which is helpful in detecting construct-irrelevant factors that may undermine the assessment validity. Moreover, listening to the test-taker's voice provides meaningful and useful information for designing an assessment that makes sense to them, especially for assessments like the one in this thesis that are unfamiliar to the test-takers and lack guidance to the test designer.

Transparent documentation of the development procedure and careful analysis of the resulting outcomes help establish a guideline for SAC assessments. 11 specific factors related to the 'construct map', 'items design', and 'outcome space' were found to influence the assessment quality and were addressed during the iterative process. The strategies corresponding to the 11 factors are:

- (1) Clarifying the context of sub-skills entailed in the assessed construct (section 5.3.1),
- (2) Editing language (section 5.3.2),
- (3) Changing scenario arrangement (section 5.3.2 and 5.4),
- (4) Balancing test length (section 5.4),
- (5) Providing basic information about SA (section 5.3.2),
- (6) Making the problem to be argued explicit (section 5.4),
- (7) Resolving item dependence (section 5.4),

- (8) Clarifying SA-related terms (section 5.4),
- (9) Considering the information provided in each task (section 5.4.2),
- (10) Changing item format (section 5.5),
- (11) Reducing the grain size of scoring rubrics (section 5.3.3).

### **10.1.2 RQ2. To what extent is the developed SAC assessment valid and reliable for assessing SAC?**

The iterative development process allows the assessment to be validated from both a micro and macro perspective, and the documentation of validity arguments shows that SAC can be measured from the proposed three components (i.e., Identifying scientific argument, Evaluating scientific argument, and Producing scientific argument) and using the SAC test results. Specifically, the process for developing the assessment is appropriately justified, with a potential weakness that the administration process was not well controlled due to the pandemic (see section 6.2.2). The assessment results are consistent with the Rasch model, with modifications for four low-performing items, and nearly no negative feedback on the assessment was reported by students (see section 6.3 and 6.4). In addition, the assessment is made transparent and easy to adapt and apply by using various data sources obtained from the assessment development process/product to develop validity arguments.

### **10.1.3 RQ3. What does the developed SAC assessment provide in terms of extended understanding of SA and of Chinese high school students' SAC?**

The complexity of the three SAC components, namely, Identification, Evaluation, and Production of SA, does not increase in a precisely linear manner. Instead, the complexity of the SAC elements within each component forms a three-level learning progression for SAC (see section 6.5). Specifically, *identifying* reason or evidence are more demanding than identifying claim and rebuttal; *evaluating* SA elements in a context that SA elements have complex connections or need more social attention is more complex; and *producing* simple SA is easy for the students as expected, thus the Production of SA crosses the three levels from generating simple SA to coherent and accurate SA. Moreover, approaching SA from a competence perspective makes it possible to integrate various SA analysis frameworks, and exploring SA as a learning progression makes different studies comparable and provides more feasible instructions for SA.

Based on the assessment and the resulted learning progression, most of the group of Chinese high school students are currently in lower levels of SAC. Specifically, 72% of the students

scored in a range that is categorized as being at level 1, while only 19.9% of the participants were at level 2. A small portion of participants were at level 0 (7.8%), and only 0.3% of the sample were at level 3.

#### **10.1.4 RQ4. How does the SAC of Chinese high school students as measured by the SAC assessment differ between different student groups?**

To explore students' SAC based on their context of the study, students in Shenzhen had significantly better performance than students in Jilin province. Students in school 1 showed significantly better performance than students in all other schools, while students in school 7 showed significantly poorer performance than students in all other schools (see section 7.2). This is almost consistent with school achievement, that is, school 1 is a school with better student achievements, while school 7 is a school with relatively poor student achievements. However, given school 1 was in Shenzhen and school 7 was in Jilin, so the SAC difference between Jilin and Shenzhen cannot be generalized as only a few schools in both areas were invited to the study. In addition, it turns out that students in key classes (i.e., with better student achievements) tend to have significantly better performance than those in ordinary classes across schools.

Providing students with the meaning of SA does not improve their SAC performance and gender does not seem to influence SAC performance (see section 7.2.4 and 7.3). Both students' school Physics test scores and school Chinese test scores have weak positive relationships with the students' SAC performance, with a stronger correlation with Chinese scores (see section 7.4).

#### **10.1.5 RQ5. What are Chinese high school students' perceptions of SA and the challenges they face in SA engagement?**

Although the cohort of students had never been invited to engage in SA in their science classrooms, they have existing understanding of SA and are able to engage in SA activities. The students are able to recognize the value of SA either generally or with a deeper understanding of how it facilitates science learning. However, their attitude toward SA seems to be twofold, namely, they are more likely to engage in SA actively when they see both the intrinsic and practical value of SA. The challenges students face in engaging in SA are also twofold. The current examination leads to exam-focused teaching and learning that does not include SA, thus students are not provided with the opportunity to engage in SA. Due to the lack of exposure to SA and the undesired impact of examination on students' learning

experience, they have developed limited competences in science learning. As a result, students have particular weaknesses in understanding/applying the social aspects compared to cognitive aspects of SA, and similarly in evaluating knowledge compared to acquiring knowledge.

## **10.2 Implications**

### **10.2.1 Implications for policy**

This section will provide implications for policy in terms of the implementation of examinations and Curriculum design. The findings of the study imply an inconsistency between what is advocated in the Curriculum documents and what is emphasized in science classrooms. In high schools in China, the Gaokao seems more powerful than the Curriculum documents in leading teaching and learning. Although there have been many discussions about whether the Gaokao should be abolished and should not be used as the only way to determine a student's admission to university, it is beyond the scope of this study to discuss whether and in what way the Gaokao should be changed. But this research argues that given the importance of Gaokao for high school students, teachers, and parents in China, it should not only serve as a tool to select students, but it should also reflect what ability/knowledge the society expects students to have. Or else, as revealed by this study, the skills that are taught and learnt in high schools will keep being narrow and unaligned with students' long-term developmental needs. Thus, putting aside the format of Gaokao, the content of Gaokao needs to change for better practice of teaching and learning, or at least consistent with what has been advocated in the Curriculum.

To design assessments including SA, although the items in the current study may not be used directly in examinations, the eleven strategies proposed in this study may be considered by the designers of school assessments or high-stakes examinations. Moreover, the proposed three-components framework and three-level learning progression for SAC may be considered when designing assessment items and deciding their complexity. The Gaokao test items have high demanding on students' understanding and reasoning ability, but students tend to view it as based on memorization. Nevertheless, students tend to view the SAC assessment as based on thinking ability. So, there is a need to include SA or other key competences in assessments in an explicit way that foreground the key features of these competences. By doing so, key competences like SA may be better understood by teachers and students and thus be better addressed in classrooms. The current study uses pencil and paper test which is limited in

capturing the social aspects of SA, therefore, other forms of assessment that can better detect students' SA engagement may be considered by assessment designers.

One of the rationales of conducting this study is the emphasis of SA in the science/Physics Curricula. The hybrid nature of SA revealed in this study suggests the need to expand and specify SA-related content in the Curriculum document. It has been well known for the policy makers of the Curriculum and the Gaokao to pay attention to thinking ability and key competences. As has been said in section 2.1, the current Curriculum includes SA as one aspect of 'scientific thinking', but there is no explicit definition of SA in the Curriculum and the interpretation for SA is far from comprehensive and specific in the Curriculum. This study revealed that SA is a complex epistemic scientific practice that needs multi-dimensional competences, adding the term SA into the Curriculum while not giving a detailed explanation this would lead to teaching practitioners either misunderstanding it or ignoring it. Therefore, this research argues that in order for SA to be emphasized in science teaching and learning, SA should be illustrated separately and by itself in the Curriculum. Namely, SA should be explained explicitly in the Curriculum, so as to make its teaching and learning explicit.

In addition, although SA is included in the category of 'scientific thinking', a closer look of the description of 'scientific inquiry' category suggests SA related procedures. As Allchin and Zemlén (2020) and this study argue, there should and does exist a scope for argumentation, as well as other scientific practices for understanding the nature of science. Using vague/incomprehensive descriptions for each unique scientific practice/competence in the curricula without discussing the differences and connections between them or between the educational merits expected from them can confuse the teachers and thus their practices in the classroom. Thus, Curriculum documents should make it clear what kind of thinking or key competences are needed, and how they are different and connected with each other to support the whole science practice and knowledge generation. By doing so, teachers would be clear about *what* they are expected to teach and thus transfer this to the students with the information of the epistemic process of science.

The Curriculum document should also include possible SA teaching and assessment strategies to inform teachers about *how* they could teach SA. Section 2.1 has discussed that the progression proposed in the curricula are not systematic and the high school curriculum provides even less comprehensive description about SA than the curricula for compulsory education. The lack of guidance for the instruction and assessment of SA and other scientific

practices/key competences in the Curricula is consistent with the lack of illustration of them. Section 9.2.2 has problematised that assessing SA in high-stakes examination may also lead to inflexible teaching and learning, therefore assessing SA is only part of the endeavour of integrating SA into classrooms. Considering the conspicuous underemphasis of SA as a pedagogy in science classrooms in China, follow up policy may be needed to illuminate how to teach and assess SA to support teachers thus to facilitate its integration into classrooms.

### **10.2.2 Implications for science teaching**

The findings of this study provide fruitful implications for teachers' practice. Most importantly, as has been recognized by previous studies (Osborne et al., 2004; Cikmaz et al., 2021), SA needs to be taught explicitly. Given SA is a cognitive, epistemic, and social practice, it is not sufficient to only rely on content knowledge to be an 'excellent' participant in SA. Explicit teaching of the epistemic knowledge and skills that are required for SA engagement will help students to gain a better understanding about the practice. Specifically, not only the meaning of SA, but the standard of high-quality SA and the expectation for students to be involved in SA should be taught to students explicitly. Similar to that argued by Khishfe (2020), both norms of argumentation and the nature of science should be emphasized in school education considering students' general understanding of SA and lack of practicing SA in science learning. Students' perceptions on the value and function of SA influence their performance and their willingness to engage. Telling them explicitly how SA and their science learning are intertwined and benefit from each other will facilitate their engagement with the activity. Overall, 'what is SA?', 'what kind of SA is high quality?', 'what role does SA plays in science enterprise?', and 'how to engage in SA appropriately?' should all be talked about explicitly in the classrooms. By doing so, teachers and students can engage in SA deliberately thus to help build the epistemic understanding of SA and epistemological understanding of science.

Scientific knowledge is built cognitively and socially in scientific discourse, thereby students should be provided with the chance to engage in dialogues that require the "coordination of the cognitive, epistemic, and social aspects of science" to think, act and speak like scientists (Groom, et al., 2018, p. 1266). Without engaging in the social process of building knowledge made students focus more on memorizing knowledge or acquiring knowledge without understanding the practice of science, which also led to students' weakness of understanding/applying the social aspects of SA. More broadly, not only SA, but other activities that can generate discussion, stimulate viewpoints, and achieve sharing should also

be introduced into school education to facilitate students' ability of constructing knowledge.

SA could even be included in examinations or school assessments, and teachers should be aware that **knowing** 'what is SA/good SA' and 'how to engage in SA' is not enough for students to acquire the competence of applying the knowledge. Students should be really provided with the opportunity to **do** SA, so that they can apply their understanding to generate an argument or an evaluation of an argument, and the practice can further enhance their understanding. As reflected from the findings, the culture and school environment around students has made them less aware of the collaborative way of constructing knowledge in SA. Moreover, it is more difficult for students to evaluate rebuttals that engage in other's arguments and generate a rebuttal that analyzes and weakens the opposing argument. Especially considering the Chinese culture where the authority of the teacher is greatly respected and dissent is not encouraged, teachers should create an environment where students find it safe and common to listen to, understand, analyze, and evaluate each other's idea (Lee et al., 2020; Chen & Qiao, 2020). A competitive atmosphere where winning a conversation is appreciated gives students more mental pressure to be involved in it. One characteristic of the Chinese culture is that it is not appreciated for people to speak out and to argue with other claims in a community where modesty is a traditional virtue, while people are competitive and care about winning or losing. Thus, in order to integrate SA in schools, it is a challenge and a necessity to build such an environment where students are encouraged and respected for expressing their ideas without worrying about whether they would be judged and if their idea is good enough, and without worrying about causing quarrels. At the same time, this environment makes both students and teachers appreciate the process value of constructing and enhancing knowledge collaboratively rather than the outcome value for winning over other students.

In terms of the cognitive aspect of SA, the learning progression resulting from this study reveals that it is harder for students to differentiate between reason and evidence and to evaluate and generate reason. This suggests that instead of focusing on the correct answer, science teaching should encourage students to interpret evidence and articulate how they get the conclusion given the evidence they use (Sampson et al., 2013). In addition, teachers should expose students to a variety of data sources and provide them with the opportunity to compare and decide which is the data that best answers the question, rather than listing all the required data for them to use directly.

Given teaching students how to construct a high-quality scientific argument requires teaching

them what counts as high quality argumentation and what counts as a rational construction of scientific knowledge, it is equally important to treat SA as a learning tool and as a learning outcome (Rapanta & Macagno, 2016). Combining the fact that students tend to have a better experience when they are dealing with authentic problems and they have existing knowledge about SA, providing them with the chance to argue about close to life issues is a good entry point to help them build the norm of SA and gain deeper understanding toward related content knowledge. Although the needs of students to be taught and assessed in an authentic context while caring about students' thinking ability has been put forward for several decades, its implementation in practice is far less often occurring (Cumming & Maxwell, 1999). Compared to continuing to ask students to do isolated drill exercise and to figure out 'far-away' questions that are exam-specific, engaging in authentic activities can improve students' ability of talking science (Lemke, 1990), and let them see its long-term connection to what they value (Baker & O'Neil, 1994).

Nevertheless, considering real life problem tends to be more complex, students' argumentation at the beginning does not need to be perfect and high-quality, but does facilitate their interest in discussion. As they progress deeper with SA and obtain more content knowledge, more sophisticated and rational SA about abstract content in the Physics curriculum such as Electromagnetism can be constructed based on the nascent form of SA. Thus, this research suggests that SA can be included in teaching and learning at the early stage of learning a module of content knowledge in authentic contexts to inspire discussion and facilitate their understanding of content knowledge, and the proficiency of content knowledge would further advance their evaluation and construction of more complex arguments. Overall, it is equally important to learning from SA and learning to conduct SA.

### **10.3 Contributions**

This study contributes to both knowledge and methodology relating to understanding SA and its assessment in the context of China for high school students. Assessments that consider multiple aspects of SA and can be used in large scale studies are still underexplored in the field of science education research. Studies that explore Chinese high school students' SA are still rare. This research provides rich insights for understanding SA from a competence perspective, exploring it as learning progression(s), developing SA assessments, and understanding Chinese high school students' experience, perception, and performance on SA. By conducting this study and discussing the findings, this research contributes a more comprehensive and feasible way



of assessing and understanding SA, highlighting the students' epistemic understanding of SA and the competences needed for SA engagement.

This study also contributes to the methodology of conducting mixed-methods research to assess SA, highlighting the contribution of the iterative assessment development design, the construction of validity arguments from both a micro and macro perspective, and the inclusion of test-takers' voices into developing an assessment.

### **10.3.1 Contribution to knowledge**

This study provides insights of understanding SA as the competences of Identifying, Evaluating, and Producing an argument, and of the potential of involving other competences into it. As discussed earlier, various studies on SA have enriched our understanding about argumentation in science education, whilst highlighting the challenge of how to make these studies comparable and how to choose between them for explicit guidance for science teaching (Henderson et al., 2018; Quinlan, 2020). These precursor studies advocate the need for a more unified way of understanding SA (Rapanta et al., 2013; Osborne et al., 2016), and thus this study has framed students' SA performance through a series of competences needed for engaging in SA. The three-components framework introduced was tested via the PCM model and the instrument constructed was found to produce a univariate measure of SA. This indicates that the competences of *Identifying*, *Evaluating*, and *Producing* a scientific argument are related components of SAC. Therefore, approaching SA from a competence perspective as this study did advances the endeavor of integrating various frameworks that put different emphasis on SA (see section 3.3) for a comprehensive investigation of SA. Importantly, this study exposes these competences to students explicitly, which takes the first step concerning the difficulty and lack of awareness students have in terms of evaluating scientific arguments, despite the fact that its importance has been underscored by many researchers (Chen et al., 2019; Tseng et al., 2021). In addition, this study also furnishes insights of expanding the SAC model by involving not only the competences that manifest directly from behaviors, but also competences that are supporting these behaviors (e.g., language use, social skill, perception, content knowledge, awareness, willingness etc.). Thus, future studies conducting SA assessment could consider using the potential SAC model introduced here.

This study also provides insights for a potential SAC learning progression by comparing it with those that have been proposed in previous studies, which promote the advocacy of making

various studies on SA comparable. In addition, it is argued that focusing on the competences needed for argumentation explicitly also provides more feasible instruction for teachers compared with complex analytical frameworks that are harder to apply in (teaching) practice. If the assessment of argumentation focuses more on the arguments students generate, from the simplest argument to the most complicated one, it is hard to decide where the end point of the continuum is, for there will always be more complex arguments possible by applying more strategies and building more connections between elements based on the problem to be argued. As a starting point for school education to cultivate students' competence of argumentation, the three-components framework and the learning progression resulted from this study serve as reliable tools to inform SA instruction in Chinese high schools and maybe in other contexts as well.

Moreover, this study contributes a validated assessment instrument for SAC, which contains not only open-ended items (that have previously usually been used in analyzing argumentation) but also multiple-choice questions that are more feasible to use in a large-scale study and contains not only construction of SA but also epistemic understanding of SA. Osborne et al. (2016) reported in their large-scale study of SA assessment that the multiple-choice items used usually underperformed and presented extremely low difficulty estimates, so they included only open-ended items in their final assessment. But they also advocated the importance of including selected response items for future study to make it applicable to large scale testing. Thus, this study advances the large-scale assessment of SAC by revealing the possibility of including such items. The previous studies on SA usually assess/analyze students' epistemic understanding of SA by analyzing their rebuttal/critique of others' arguments, while this instrument provides ideas on letting students evaluate SA directly with explicit criteria. The three-components assessment framework, the learning progression, and the assessment instrument together expand the conceptualization and assessment of SA and so offer a unique perspective that can inform the design of SA assessments, especially for the written form of SA.

Lastly, this study provides empirical evidence to understand how Chinese high school students experience, perceive and perform SA. As mentioned in section 3.4.3, there is a conspicuous lack of studies exploring SA from students' perspectives especially in the context of China. However, students' perception of SA is significant for implementing SA, and listening to their voices provides more direct and authentic information in terms of the challenges they face in

SA engagement. This study also found that, similar to studies conducted in other countries, Chinese students in general appreciate their experience of engaging in SA. Thus, this study adds to the literature in terms of the performance and weaknesses of Chinese high school students' SA, and their perceptions of SA. By understanding Chinese high school students' SA, this study provides information that helps pinpoint the appropriate ways and focuses of the teaching and learning of SA.

### **10.3.2 Contribution to methodology**

This study provides insights for the methodological approaches in investigating the assessment and understanding of SA. Firstly, this study expands on assessment development methodologies of conducting an iterative assessment development process (Wilson, 2004; AERA, 2018), by applying the findings from the iterative procedure to support developing the assessment and documenting these findings as an approach for developing a SAC assessment. This method does not only make the process of assessment development transparent, but it also contributes guidelines for future SA assessment development. In this way, not only an instrument itself and the findings found by using the instrument matter for the research community, the design of the instrument can also be comparable across the research field thus to be a continuous endeavor. This is especially important in areas such as SA where assessments are under-explored or where assessments are plagued by too many frameworks. In addition, the assessment instrument in this study is not only a tool for data collection, but it also serves as a platform for me as a researcher to talk with the students about SA and as a pedagogical resource for the students to think, learn, and reflect on their understanding of SA.

This study also contributes to the practice of validity theories in the field of SA assessment. Specifically, this study combined Newton's (2017) macro and micro validation theory and Kane's (2013) argument-based validation theory and put them into practice for the benefit of validating a SA assessment. By doing so, both the researcher and the potential reader of this study can better examine the product and the process of developing the assessment, as well as its administration. Similar to an iterative design method, formulating a validation argument from both macro and micro aspects enabled me as a researcher to not only demonstrate the findings but also to evaluate the assessment by myself and reflect on it.

Lastly, this study furnishes insights about involving the students/test-takers' voices when investigating a construct and its assessment and adds to literature about students' perspectives

on SA and its assessment. In the area of assessment development, it has been often the case that emphasis is on investigating professionals/teachers' ideas about the assessment, although including think aloud and follow-up interviews have been advocated (Wilson, 2004; AERA, 2018). This study suggests how these approaches contribute to a broader understanding of the SA construct and its assessment by carefully analyzing the interviews with students, especially students in high schools who are cognitively mature, to provide fruitful information.

Overall, this study demonstrates how these methods are relevant and can benefit the research about assessing and understanding SA, therefore enlarging the methodologies that can be used to conduct research on SA.

#### **10.4 Limitations and future research**

This study has provided rich information in terms of the nature of SA, the way to assess SA, and the students' understanding of SA. I acknowledge the limitations of this study in the construction of SAC, the design of the assessment instrument, and the representativeness of the results due to the limited knowledge of myself as a researcher, the limited resources and time of doing the research, and the influence of the COVID-19 pandemic. Therefore, this section will discuss the limitations of this study and propose possible future studies in the interest of potential readers of this study (e.g., science teachers, science education researchers).

Firstly, considering the feasibility, only students in their second year of high school (Year 11) were recruited in this study. However, involving students in Year 10 and Year 12 would capture a more comprehensive picture of how Chinese high school students understand SA, and would better examine whether and how students in different school years progress along the SAC learning progression. In particular considering students in their last year of high school usually do not learn new knowledge instead they review the previous knowledge they had learnt and thus have more enhanced content knowledge. Longitudinal studies would better track and examine students' development of SAC and thus check the learning progression of SA. In addition, this study only included schools in urban areas in eastern China, so involving schools in rural area or in western China would provide richer information on students' performance across areas.

Secondly, the SAC assessment only includes topics related to 'motion and force' to make sure all the participants had learnt the content knowledge. Assessments that involve other topics and more complex content knowledge need to be developed to facilitate the use of SAC

assessments in schools.

Thirdly, due to the lack of resource, this study didn't involve professionals in SA to review the assessment instrument, although high school science teachers and science education researchers were invited. As there are few researchers in China investigating SA, involving more professionals in SA to discuss the assessment instrument would generate more creative ideas on how to design it and thus would potentially lead to a higher quality assessment. Moreover, items Erb\_7.2 and Ee\_7.3 showed underfit with the PCM and some P-SA items showed local dependence. Such items need further exploration in future studies.

Additionally, the SAC assessment did not capture the dynamic argumentation process, in which students may show different acts related or unrelated to argumentation. As previously mentioned, this study considered the pedagogical function of the assessment when designing it as helping students to learn SA and its procedure. However, going through the process of SA should also be the means rather than only the ends (Berland et al., 2016), whereas assessments that focus on the social dimension of SA can help students to know how to learn *from* SA (Clark et al., 2007). Although it has been argued that written argumentation is important for science learning (Osborne et al., 2016) and this study argues that written argumentation can be taken as a starting point of transforming the science classroom to be more open and collaborative where SA is embedded comfortably, it is also necessary to explore ways of extending SAC and its progression levels to the social dimension of SA. Moreover, further exploring students' experience of engaging in group argumentation can help researchers better understand the similarities and differences of students' experience when engaging in written form SA and social form SA, so to further implicate the assessment and teaching in these two forms.

Fifthly, this study did not ask students to generate evaluation directly considering students were not familiar with this practice. Investigating how students generate evaluation for arguments would provide more insights on students' epistemic understanding of SA and understanding of the nature of science. Thus, further study could consider expanding the E-SA component by asking students to produce their evaluation of an argument or competing arguments with or without provided criterion.

Furthermore, SA is an 'entity' that is more complex than stacking up a set of sub-skills or knowledge items. The findings of this study not only show the challenges students face in

engaging in SA but also uncover what a complex task it is to assess students' SAC especially with a pencil and paper test. Thus, careful consideration needs to be taken as to how to incorporate or balance the focus of SA assessment as to the essential constitutes and sophistication of argumentation, the understanding of scientific knowledge, the rhetorical effectiveness, and the goal achievement of the activity (such as persuasiveness). Compared to 'what is scientific argumentation' and 'what is the content knowledge', how to include 'what counts as good (argumentation or scientific practice)', 'how to engage in SA socially' and 'how to interact effectively' into assessment needs more research. In addition, assessment tasks that provide open solutions and allow deep thinking and complex argumentation in Physics are needed.

Lastly, this study only explored the students' experience, understanding and performance on SA. It would also be valuable to explore how teachers perceive SA, its assessment, and its implementation in classrooms, and to investigate the pedagogical skills they have already used and that is possible to facilitate teaching practices for talking and doing SA.

To sum up, despite the contributions made by this study, it raises further questions for exploring the nature and the assessment of SA. More empirical evidence is needed for an in-depth understanding of the construct of SAC and assessing it in a comprehensive, feasible, and comparable way, and for richer information about how SA can be facilitated in school education in China and other contexts. Thus, this study appeals for more research to be conducted to explore the construct of SAC and to continue uncovering more markers along the underlying continuum of it, and more research conducted in China and across the world to promote its implementation in science classrooms.

In retrospect, doing research on scientific argumentation has been a challenge in itself given its connection to fields such as Philosophy, Science, Logic, Psychology etc. While performing the assessment of SA poses further challenges since it also needs expertise in assessment and measurement. It has been promising that this study justified understanding SA as a series of competences and learning progression (s), generated guidelines for assessing SA by employing an iterative approach, and proposed validity arguments focusing on both the process and product of assessment. However, what has impressed me most is that the participants' experience of their school learning is so similar to mine as mentioned in section 1.1.3, although over ten years have passed. Despite the implicit or explicit emphasis in the curricula that schooling should go beyond content knowledge, opportunities for students to engage in

activities such as SA remain limited. It's surprising that the participants expressed positive attitudes towards SA. So, now is time to consider seriously in terms of how to integrate SA into school education. Assessing SA is an indispensable part of this endeavour.

## **COVID-19 statements**

The influence of COVID-19 on conducting the study has been mentioned in the thesis. Overall, there are two main impacts, namely, delays of collecting data and inaccessibility to classrooms.

The pandemic made my PhD journey difficult for a long time, especially when the data collection schedule was delayed for around 6 months. The time left for me to analyse data and writing up was tight. Additionally, as mentioned in section 4.4.2, the test papers were handed over to teachers to manage, therefore to a certain extent, caused inconsistencies in the operation between schools. If it's not because of the pandemic, I could enter the classrooms to collect test data to ensure that the test administration is consistent across schools and to know more about what might be happening in the classrooms. The impact of unable to enter classrooms may have resulted in relatively lower-quality test data compared to without the pandemic.

In addition, all the interview with the students were conducted via voice calls. Without pandemic, I could have interviewed the students in person so that more information in terms of their body language or facial expression can be obtained. However, this impact, while ruling out other possibilities, does not seem to influence the quality of the study given some students mentioned that they would be nervous if they were interviewed in person.



## Reference

- Abu-Alhija, F. N. (2007). Large-Scale Testing: Benefits and Pitfalls. *Studies in Educational Evaluation*, 33(1), 50–68.
- Ahmed, A., & Pollitt, A. (2000, May). Observing context in action. In International Association of Educational Assessment Conference (IAEA), Jerusalem, Israel.
- Ahmed, A., & Pollitt, A. (2001, September). Improving the validity of contextualised questions. In British Educational Research Association Conference, Leeds.
- Akyel, A., & Kamisli, S. (1996). Composing in First and Second Languages: Possible Effects of EFL Writing Instruction.
- Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational research*, 58(4), 375-404.
- Allchin, D., & Zemplén, G. Á. (2020). Finding the place of argumentation in science education: Epistemics and Whole Science. *Science Education*, 104(5), 907-933.
- Amano, I., & Poole, G. S. (2005). The Japanese university in crisis. *Higher Education*, 50(4), 685–711.
- American Educational Research Association. (2018). Standards for educational and psychological testing. American Educational Research Association.
- An, X., & Yung, Y. F. (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. SAS Institute Inc. SAS364-2014, 10(4).
- Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback—a case-study. *System*, 30(2), 207-223.
- Anderson, L.W., & Krathwohl (Eds.). (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.
- Angell, R. B. (1964). Reasoning and logic.
- Archibald, D. A., & Newman, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in secondary schools*. Washington, DC: National Association of Secondary School Principals.
- Australian Curriculum and Assessment Reporting Authority [ACARA], *The Australian Curriculum: Science*, 2016.
- Bai, C., Chi, W., & Qian, X. (2014). Do college entrance examination scores predict undergraduate GPAs? A tale of two universities. *China Economic Review*, 30, 632-674.
- Baird, J. A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart?. *Assessment in Education: Principles, Policy & Practice*, 24(3), 317-350.
- Baker, E. L., & O'Neil Jr, H. F. (1994). Performance assessment and equity: A view from the USA. *Assessment in Education: principles, policy & practice*, 1(1), 11-26.
- Bathgate, M., Crowell, A., Schunn, C., Cannady, M., & Dorph, R. (2015). The learning benefits of being willing and able to engage in scientific argumentation. *International Journal of Science Education*, 37(10), 1590–1612.
- Becker, C. B. (1986). Reasons for the lack of argumentation and debate in the Far East. *International Journal of Intercultural Relations*, 10(1), 75–92.
- Berland, L. K., & Hammer, D. (2012). Students' framings and their participation in scientific argumentation. In *Perspectives on scientific argumentation* (pp. 73-93). Springer, Dordrecht.
- Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, 94(5), 765-793.
- Berland, L. K., & McNeill, K. L. (2012). For Whom Is Argument and Explanation a Necessary Distinction? A Response to Osborne and Patterson. *Science Education*, 96(5), 808-813.

- Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science education*, 93(1), 26-55.
- Berland, L. K., & Reiser, B. J. (2011). Classroom communities' adaptations of the practice of scientific argumentation. *Science Education*, 95(2), 191-216.
- Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2016). Epistemologies in practice: Making scientific practices meaningful for students. *Journal of Research in Science Teaching*, 53(7), 1082-1112.
- Biesta, G. J. J., & Burbules, N. C. (2003). *Pragmatism and Educational Research*. (Philosophy, Theory, and Educational Research Series). Rowman & Littlefield.
- Biggs, J. B., and K. F. Collis. 1982. *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. New York: Academic Press
- Billig, M., Condor, S., Edwards, D., Gane, M., Middleton, D., & Radley, A. (1988). *Ideological dilemmas: A social psychology of everyday thinking*: Sage Publications, Inc.
- Bishop, N. S., & Davis-Becker, S. (2015). Preparing examinees for test taking: Guidelines for test developers. In *Handbook of test development* (pp. 570-582). Routledge.
- Blair, J. A. (2012). The rhetoric of visual arguments. In *Defining visual rhetorics* (pp. 53-74). Routledge.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), 3–13.
- Bond, T., & Fox, C.M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, Third Edition (3rd ed.). Routledge.
- Bowen, T. (2017). Assessing visual literacy: a case study of developing a rubric for identifying and applying criteria to undergraduate student learning. *Teaching in Higher Education*, 22(6), 705-719.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77-101.
- Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health*, 11(4), 589-597.
- Bricker, L. A., & Bell, P., (2007). Evidentiality and evidence use in children's talk across everyday contexts. Paper presented at the annual meeting of the Society for Social Studies of Science (4S), Montreal, Canada.
- Bricker, L. A., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. *Science Education*, 92, 473–498.
- Bricker, L. A., & Bell, P. (2012). Argumentation and reasoning in life and in school: Implications for the design of school science learning environments. In *Perspectives on scientific argumentation* (pp. 117-133). Springer, Dordrecht.
- Brigandt, I. (2016). Why the difference between explanation and argument matters to science education. *Science & Education*, 25(3), 251-275.
- Britt, M. A., Richter, T., & Rouet, J. F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist*, 49(2), 104-122.
- Browne, W. J., & Rasbash, J. (2011). *What is multilevel modelling?*. Sage.
- Burner, T. (2014). The potential formative benefits of portfolio assessment in second and foreign language writing contexts: A review of the literature. *Studies in Educational Evaluation*, 43, 139–149.
- Butler, Y. G., Peng, X., & Lee, J. (2021). Young learners' voices: Towards a learner-centered approach to understanding language assessment literacy. *Language Testing*, 38(3), 429-455.
- Cavagnetto, A. R. (2010). *Argument to foster scientific literacy: A review of argument*

- interventions in K–12 science contexts. *Review of Educational Research*, 80(3), 336-371.
- Cavagnetto, A., Hand, B. M., & Norton-Meier, L. (2010). The nature of elementary student science discourse in the context of the science writing heuristic approach. *International Journal of Science Education*, 32(4), 427-449.
- Cetin, P. S. (2014). Explicit argumentation instruction to facilitate conceptual understanding and argumentation skills. *Research in Science & Technological Education*, 32(1), 1-20.
- Charters, E. (2003). The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education Journal*, 12(2).
- Chen, Y. C., Aguirre-Mendez, C., & Terada, T. (2020). Argumentative writing as a tool to develop conceptual and epistemic knowledge in a college chemistry course designed for non-science majors. *International Journal of Science Education*, 42(17), 2842-2875.
- Chen, Y. C., Benus, M. J., & Hernandez, J. (2019). Managing uncertainty in scientific argumentation. *Science Education*, 103(5), 1235-1276.
- Chen, Y. C., & Terada, T. (2021). Development and validation of an observation-based protocol to measure the eight scientific practices of the next generation science standards in K-12 science classrooms. *Journal of Research in Science Teaching*, 58(10), 1489-1526.
- Chen, Y. C., & Qiao, X. (2020). Using students' epistemic uncertainty as a pedagogical resource to develop knowledge in argumentation. *International Journal of Science Education*, 42(13), 2145-2180.
- Chen, T., Kung, J. K. S., & Ma, C. (2020). Long live Keju! The persistent effects of China's civil examination system. *The economic journal*, 130(631), 2030-2064.
- Cheng, L. (2000). A review of the impact of testing on teaching and learning. *Non-Journal - Opinion Paper*, 34.
- Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing*, 28(2), 221-249.
- Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. *Washback in language testing*, 25-40.
- Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational assessment*, 16(2), 104-122.
- Cikmaz, A., Fulmer, G., Yaman, F., & Hand, B. (2021). Examining the interdependence in the growth of students' language and argument competencies in replicative and generative learning environments. *Journal of Research in Science Teaching*.
- Clark, D. B., & Sampson, V. D. (2005). Analyzing the quality of argumentation supported by personally-seeded discussions.
- Clark, D. B., Sampson, V., Weinberger, A., & Erkens, G. (2007). Analytic frameworks for assessing dialogic argumentation in online learning environments. *Educational Psychology Review*, 19(3), 343-374.
- Cockerill, T. (1989). The kind of competence for rapid change. *Personnel management*, 21(9), 52-56.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.
- Cohen, L., Manion, L., & Morrison, K. (2002). *Research methods in education*. routledge.
- Cohen, A. D., & Upton, T. A. (2006). Strategies in responding to the new TOEFL reading tasks. *ETS Research Report Series*, 2006(1), i-162.
- Crisp, V., Sweiry, E., Ahmed, A., & Pollitt, A. (2008). Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions. *Educational Research*, 50(1), 95-115.
- Cui, Y. (2001). Guojia kecheng biao zhun yu kuangjia de jiedu [Interpretation of National

- Curriculum Standards and Frameworks]. *Quanqiu jiaoyu zhanwang*, (08), 4-9.
- Cumming, J., & Maxwell, G. S. (1999). Contextualising authentic assessment. *Assessment in education: Principles, policy & practice*, 6(2), 177-194.
- Dai, L., Xu, P. (2021). *Jiyu qingjing leixingxue de gaokao wuli shiti qingjing de bijiao yanjiu-yi 2021 nian ge shengshi gaokao wuli shiti weili* [A comparative study of the college entrance examination Physics test questions based on situation typology—Taking the college entrance examination physics test questions of various provinces and cities in 2021 as an example]. *Wuli jiaoshi*, (12), 140
- Dawson, V., & Carson, K. (2017). Using climate change scenarios to assess high school students' argumentation skills. *Research in Science & Technological Education*, 35(1), 1-16.
- Deane, P., Song, Y., van Rijn, P., O'Reilly, T., Fowles, M., Bennett, R., ... & Zhang, M. (2019). The case for scenario-based assessment of written argumentation. *Reading and Writing*, 32(6), 1575-1606.
- Deng. (2015). *The research on the assessment scientific argumentation*. Central China Normal University.
- Deng, & Wang. (2014). The necessity of infuse scientific argumentation into science education-based on the perspective of science, science learning and international comparasion. *Elementary & Secondary Schooling Abroad*(3), 60-65.
- Deng, Y., & Wang, H. (2017). Research on evaluation of Chinese students' competence in written scientific argumentation in the context of chemistry. *Chemistry Education Research and Practice*, 18(1), 127-150.
- Department for Education [DfE], *The national curriculum in England: Framework document*, 2014.
- Dewey, J. (1929). *The quest for certainty*. Minton, Balch.
- Dewey, J. (1938). *Logic: the theory of inquiry*. Holt.
- Dewey, J. (1939). *Experience, knowledge and value: A rejoinder*. na.
- Dewey, J. (1949). *Experience and existence: A comment*. *Philosophy and Phenomenological Research*, 9(4), 709-713.
- DiBattista, D., Sinnige-Egger, J. A., & Fortuna, G. (2014). The “none of the above” option in multiple-choice testing: An experimental study. *The Journal of Experimental Education*, 82(2), 168-183.
- Doyle, L., Brady, A. M., & Byrne, G. (2009). An overview of mixed methods research. *Journal of research in nursing*, 14(2), 175-185.
- Dong. (2008). Collaborative Reasoning in China and Korea. *Reading Research Quarterly*, 43(4), 400–424.
- Dong. (2018). The new progress of international science education in 21st century-based on CiteSpace analysis. *Modern Education Management*(5), 98-105.
- Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science education*, 84(3), 287-312.
- Duncan, R. G., & Hmelo-Silver, C. E. (2009). *Learning progressions: Aligning curriculum, instruction, and assessment*.
- Duncan, P. W., Bode, R. K., Lai, S. M., Perera, S., & Glycine Antagonist in Neuroprotection Americas Investigators. (2003). Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. *Archives of physical medicine and rehabilitation*, 84(7), 950-963.
- Dunn, K. (2005) ‘Interviewing’, in I. Hay (ed.) *Qualitative Research Methods in Human Geography* (2nd edn). Melbourne: Oxford University Press, pp. 79–105.
- Duschl, R. (2008). *Science Education in Three-Part Harmony: Balancing Conceptual,*

- Epistemic, and Social Learning Goals. *Review of Research in Education*, 32(1), 268–291.
- Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education.
- Eccles, D. W., & Arsal, G. (2017). The think aloud method: what is it and how do I use it?. *Qualitative Research in Sport, Exercise and Health*, 9(4), 514-531.
- Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: perceived importance, invested effort, and test anxiety. *European journal of psychology of education*, 28(2), 497-510.
- Elwood, J., Hopfenbeck, T., & Baird, J. A. (2017). Predictability in high-stakes examinations: Students' perspectives on a perennial assessment dilemma. *Research Papers in Education*, 32(1), 1-17.
- Emler, T. E., Zhao, Y., Deng, J., Yin, D., & Wang, Y. (2019). Side effects of large-scale assessments in education. *ECNU Review of Education*, 2(3), 279-296.
- Erduran, S., Ozdem, Y., & Park, J. Y. (2015). Research trends on argumentation in science education: A journal content analysis from 1998–2014. *International Journal of STEM Education*, 2(1), 1-12.
- Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science education*, 88(6), 915-933.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological review*, 87(3).
- Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1), 1-4.
- Evagorou, M., & Osborne, J. (2013). Exploring young students' collaborative argumentation within a socioscientific issue. *Journal of Research in Science Teaching*, 50(2), 209-237.
- Felton, M., Garcia-Mila, M., & Gilabert, S. (2009). Deliberation versus dispute: The impact of argumentative discourse goals on learning and reasoning in the science classroom. *Informal Logic*, 29(4), 417-446.
- Felton, M., Garcia-Mila, M., Villarroel, C., & Gilabert, S. (2015). Arguing collaboratively: Argumentative discourse types and their potential for knowledge building. *British Journal of Educational Psychology*, 85(3), 372–386.
- Ferrara, S., & Lai, E. (2015). Documentation to support test score interpretation and use. In *Handbook of test development* (pp. 619-639). Routledge.
- Fischer, F., Wecker, C., Hetmanek, A., Osborne, J., Chinn, C. A., Duncan, R. G., . . . Sandoval, W. A. (2014). The interplay of domain-specific and domain-general factors in scientific reasoning and argumentation. In: Boulder, CO: International Society of the Learning Sciences.
- Ford, M. J. (2012). A dialogic account of sense-making in scientific argumentation and reasoning. *Cognition and Instruction*, 30(3), 207-245.
- Gallo, A. M., Sheehy, D. A., Patton, K., & Griffin, L. (2006). Assessment Benefits and Barriers. *Journal of Physical Education, Recreation & Dance*, 77(8), 46–50.
- Gan, C. (2002). Decline of the imperial examination system and the disintegration of institutional Confucianism. *Social Sciences in China*, 2, 107–117 (in Chinese).
- Garcia-Mila, M., & Andersen, C. (2007). Cognitive foundations of learning argumentation. In *Argumentation in science education* (pp. 29-45). Springer, Dordrecht.
- Garcia-Mila, M., Gilabert, S., Erduran, S., & Felton, M. (2013). The effect of argumentative task goal on the quality of argumentative discourse. *Science Education*, 97(4), 497-523.
- Gee, J. P., & Green, J. L. (1998). Chapter 4: Discourse analysis, learning, and social practice: A methodological study. *Review of research in education*, 23(1), 119-169.

- Gilbert, J. (2005). *Catching the knowledge wave? The knowledge society and the future of education*. Wellington, New Zealand: NZCER Press.
- Giuliodori, M. J., Lujan, H. L., & DiCarlo, S. E. (2008). Collaborative group testing benefits high- and low-performing students. *American Journal of Physiology - Advances in Physiology Education*, 32(4), 274–278.
- Glassner, A. (2017). Evaluating arguments in instruction: Theoretical and practical directions. *Thinking Skills and Creativity*, 24, 95-103.
- Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*.
- Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, 1(4), 223–231.
- Goldman, A. (1999). Knowledge in a social world.
- González-Howard, M., & McNeill, K. L. (2020). Acting with epistemic agency: Characterizing student critique during argumentation discussions. *Science Education*, 104(6), 953-982.
- Govier, T. (2018). Problems in argument analysis and evaluation. *Windsor studies in argumentation* (Book 6).
- Grooms, J., Sampson, V., & Enderle, P. (2018). How concept familiarity and experience with scientific argumentation are related to the way groups participate in an episode of argumentation. *Journal of Research in Science Teaching*, 55(9), 1264-1286.
- Guest, G. (2013). Describing mixed methods research: An alternative to typologies. *Journal of mixed methods research*, 7(2), 141-151.
- Hager, P., & Gonczi, A. (1996). What is competence?. *Medical teacher*, 18(1), 15-18.
- Halpern, J. Y., & Vardi, M. Y. (1989). The complexity of reasoning about knowledge and time. I. Lower bounds. *Journal of Computer and System Sciences*, 38(1), 195-237.
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.
- Han, Hu, & Wang. (2014). The progress and trend of International science teaching *Journal of East China Normal University (Educational Sciences)*, 32(4), 63-70.
- Handke, L., & Barthauer, L. (2019). Heider (1958): The Psychology of Interpersonal Relations. In *Schlüsselwerke der Netzwerkforschung* (pp. 259-262). Springer VS, Wiesbaden.
- Heitmann, P., Hecht, M., Scherer, R., & Schwanewedel, J. (2017). “Learning Science Is About Facts and Language Learning Is About Being Discursive”—An Empirical Investigation of Students' Disciplinary Beliefs in the Context of Argumentation. *Frontiers in psychology*, 8, 946.
- Henderson, J. B., MacPherson, A., Osborne, J., & Wild, A. (2015). Beyond construction: Five arguments for the role and value of critique in learning science. *International Journal of Science Education*, 37(10), 1668-1697.
- Henderson, J. B., McNeill, K. L., González-Howard, M., Close, K., & Evans, M. (2018). Key challenges and future directions for educational research on scientific argumentation. *Journal of Research in Science Teaching*, 55(1), 5-18.
- Hillocks, G. (2002). *The testing trap: How state writing assessments control learning*. Teachers College Press.
- Hogan, K., & Maglienti, M. (2001). Comparing the epistemological underpinnings of students' and scientists' reasoning about conclusions. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 38(6), 663-687.
- Hohensinn, C. (2018). pcIRT: an R package for polytomous and continuous Rasch models. *Journal of Statistical Software*, 84, 1-14.
- Hu, Y., Tang, Y. (2013). Gaozhongsheng kexuesuyang de xingbie chayi-jiyu wutiaojianfenweishuhuigui de jingyan yanjiu [Gender differences in scientific literacy

- among high school students: An empirical study based on unconditional quantile regression]. *Bejingdaxue jiaoyu pinglun*, (4), 110-128.
- Hudicourt-Barnes, J. (2003). The use of argumentation in Haitian Creole science classrooms. *Harvard Educational Review*, 73(1), 73–93.
- Iordanou, K. (2010). Developing argument skills across scientific and social domains. *Journal of Cognition and Development*, 11(3), 293-327.
- Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, 49(1), 82-100.
- Jiménez-Aleixandre, M. P., Bugallo Rodríguez, A., & Duschl, R. A. (2000). “Doing the lesson” or “doing science”: Argument in high school genetics. *Science education*, 84(6), 757-792.
- Jiménez-Aleixandre M.P., Erduran S. (2007) Argumentation in Science Education: An Overview. In: Erduran S., Jiménez-Aleixandre M.P. (eds) *Argumentation in Science Education*. Science & Technology Education Library, vol 35. Springer, Dordrecht.
- Jiménez-Aleixandre, M. P., & Puig, B. (2012). Argumentation, evidence evaluation and critical thinking. In *Second international handbook of science education* (pp. 1001-1015). Springer, Dordrecht.
- Jin, H., Hokayem, H., Wang, S., & Wei, X. (2016). A US-China interview study: Biology students’ argumentation and explanation about energy consumption issues. *International Journal of Science and Mathematics Education*, 14(6), 1037-1057.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, 33(7), 14-26.
- Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research*, 1(2), 112-133.
- Kane, M. (2006). Content-related validity evidence in test development. *Handbook of test development*, 1, 131-153.
- Kane, M. (2012). All validity is construct validity. Or is it?. *Measurement: Interdisciplinary Research & Perspective*, 10(1-2), 66-70.
- Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, 42(4), 448-457.
- Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198-211.
- Kaushik, V., & Walsh, C. A. (2019). Pragmatism as a research paradigm and its implications for social work research. *Social sciences*, 8(9), 255.
- Kaya, E., Erduran, S., & Cetin, P. S. (2010). High school students’ perceptions of argumentation. *Procedia-Social and Behavioral Sciences*, 2(2), 3971-3975.
- Kaya, E., Erduran, S., & Cetin, P. S. (2012). Discourse, argumentation, and science lessons: match or mismatch in high school students’ perceptions and understanding? Special Issue on Inquiry in Science Education & Argumentation Based Scientific Inquiry. *Mevlana International Journal of Education*, 2(3), 1-32.
- Ke, L., Sadler, T. D., Zangori, L., & Friedrichsen, P. J. (2020). Students’ perceptions of socio-scientific issue-based learning and their appropriation of epistemic tools for systems thinking. *International Journal of Science Education*, 42(8), 1339-1361.
- Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. *Science Education*, 86(3), 314–342.
- Kelly, G. (2008). Inquiry, activity and epistemic practice. In *Teaching scientific inquiry* (pp. 99-117): Brill Sense.
- Khishfe, R. (2014). Explicit nature of science and argumentation instruction in the context of socioscientific issues: An effect on student learning and transfer. *International Journal of Science Education*, 36(6), 974-1016.

- Khishfe, R. (2020). Retention of acquired argumentation skills and nature of science conceptions. *International Journal of Science Education*, 42(13), 2181-2204.
- Khine, M. S. (Ed.). (2011). *Perspectives on scientific argumentation: Theory, practice and research*. Springer Science & Business Media.
- Kizlik, B. (2012). *Measurement, Assessment, and Evaluation in Education*. Retrieved October, July, 1–43.
- Kohlberg, L., & Mayer, R. (1972). Development as the aim of education. *Harvard educational review*, 42(4), 449-496.
- Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. New Jersey: Prentice-Hall.
- Koepfen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie/Journal of Psychology*, 216(2), 61-73.
- Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.
- Kuang, A. (2019). *Gaozhongsheng kemu xuanze Zhong de xingbie chayi xianxiang yanjiu* [A study on gender differences in subject choice of high school students]. *Zhejiangdaxue*.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (Vol. 111). University of Chicago Press: Chicago.
- Kuhn, D. (1991). *The skills of argument*. Cambridge University Press.
- Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Sci. Ed.*, 77: 319-337.
- Kuhn, D. (2015). Thinking together and alone. *Educational Researcher*, 44(1), 46–53.
- Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive development*, 15(3), 309-328.
- Kuhn, L., & Reiser, B. (2005, April). Students constructing and defending evidence-based scientific explanations. In annual meeting of the National Association for Research in Science Teaching, Dallas, TX (pp. 1-35).
- Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child development*, 74(5), 1245-1260.
- Kuhn, D., & Udell, W. (2007). Coordinating own and other perspectives in argument. *Thinking & Reasoning*, 13(2), 90-104.
- Kuhn, D., Wang, Y., & Li, H. (2010). Why argue? Developing understanding of the purposes and values of argumentative discourse. *Discourse processes*, 48(1), 26-49.
- Kuhn, D., Zillmer, N., Crowell, A., & Zavala, J. (2013). Developing norms of argumentation: Metacognitive, epistemological, and social dimensions of developing argumentative competence. *Cognition and Instruction*, 31(4), 456-496.
- Kvale, S. (1996). The 1,000-page question. *Qualitative inquiry*, 2(3), 275-284.
- Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). *Handbook of test development* (pp. 3-18). New York, NY: Routledge.
- Larson, A. A., Britt, M. A., & Kurby, C. A. (2009). Improving students' evaluation of informal arguments. *The Journal of Experimental Education*, 77(4), 339-366.
- Le Deist, F. D., & Winterton, J. (2005). What is competence?. *Human resource development international*, 8(1), 27-46.
- Leitão, S. (2000). The potential of argument in knowledge building. *Human Development*, 43(6), 332–360.
- Lee, H. S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in*



- Science Teaching, 51(5), 581-605.
- Lee, S. W. Y., Liang, J. C., & Tsai, C. C. (2016). Do sophisticated epistemic beliefs predict meaningful learning? Findings from a structural equation model of undergraduate biology learning. *International Journal of Science Education*, 38(15), 2327-2345.
- Lee, H., Lee, H., & Zeidler, D. L. (2020). Examining tensions in the socioscientific issues classroom: Students' border crossings into a new culture of science. *Journal of Research in Science Teaching*, 57(5), 672-694.
- Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Ablex Publishing Corporation, 355 Chestnut Street, Norwood, NJ 07648 (hardback: ISBN-0-89391-565-3; paperback: ISBN-0-89391-566-1).
- Leung, J. S. C. (2020). Students' adherences to epistemic understanding in evaluating scientific claims. *Science Education*, 104(2), 164-192.
- Leutner, D., Fleischer, J., Grünkorn, J., & Klieme, E. (2017). Competence assessment in education: An introduction. In *Competence assessment in education* (pp. 1-6). Springer, Cham.
- Levy, C. M., & Ransdell, S. (2013). *The science of writing: Theories, methods, individual differences and applications*. Routledge.
- Li, J. (2006). Self in learning: Chinese adolescents' goals and sense of agency. *Child Development*, 77, 482-501.
- Li, F. (2009). Jiyu kecheng biao zhun de jiaoxue: cong "wenben kecheng" dao "jiaoxue shijian" [Teaching Based on Curriculum Standards: From "Curriculum text" to "Teaching practice"]. *Dangdai jiaoyu kexue*, 2009(16):6-9.
- Li, L., & Wang, Z. (2022). Understanding the long-term effects of Keju: The case of entrepreneurship in China. *Economics of Transition and Institutional Change*.
- Li, X., Deng, L., Zhang, X. (2022). 2021 nian xingao kao wuli shiti zhong guanjian nengli de kaocha yanjiu [Research on the key abilities in the new college entrance examination physics items in 2021]. *Wuli jiaoshi*, 43(2), 74.
- Lincoln, Y. S. (2010). "What a long, strange trip it's been...": Twenty-five years of qualitative and new paradigm research. *Qualitative inquiry*, 16(1), 3-9.
- Lincoln, Yvonne, Susan A. Lynham, and Egon G. Guba. 2011. Paradigms and perspectives in contention. In *The Sage Handbook of Qualitative Research*. Edited by Norman K. Denzin and Yvonna S. Lincoln. Thousand Oaks: Sage Publications, pp. 91-95.
- Liu, Y. (2013). Meritocracy and the Gaokao: a survey study of higher education selection and socio-economic participation in East China. *British Journal of Sociology of Education*, 34(5-6), 868-887.
- Liu, Q. T., Liu, B. W., & Lin, Y. R. (2019). The influence of prior knowledge and collaborative online learning environment on students' argumentation in descriptive and theoretical scientific concept. *International Journal of Science Education*, 41(2), 165-187.
- Liu, G. X. Y., & Helwig, C. C. (2020). Autonomy, social inequality, and support in Chinese urban and rural adolescents' reasoning about the Chinese college entrance examination (Gaokao). *Journal of Adolescent Research*, 0743558420914082.
- Liu, H., & Wu, Q. (2006). Consequences of college entrance exams in China and the reform challenges. *KEDI Journal of Educational Policy*, 3(1).
- Llewellyn, D. (2013). *Teaching high school science through inquiry and argumentation*. Corwin Press.
- Lombardi, D., Bailey, J. M., Bickel, E. S., & Burrell, S. (2018). Scaffolding scientific thinking: Students' evaluations and judgments during Earth science knowledge construction. *Contemporary Educational Psychology*, 54, 184-198.
- Longhurst, R. (2003). Semi-structured interviews and focus groups. *Key methods in geography*, 3(2), 143-156.

- Luo, A. (2005). “Xueeryouzeshi” bian [“Study and excellence lead to an official career” argument]. *Zhongguo zhexue shi*, (3), 31-38.
- Luo, X., Wei, B., Shi, M., & Xiao, X. (2020). Exploring the impact of the reasoning flow scaffold (RFS) on students’ scientific argumentation: based on the structure of observed learning outcomes (SOLO) taxonomy. *Chemistry Education Research and Practice*, 21(4), 1083-1094.
- Lytzerinou, E., & Iordanou, K. (2020). Teachers’ ability to construct arguments, but not their perceived self-efficacy of teaching, predicts their ability to evaluate arguments. *International Journal of Science Education*, 42(4), 617-634.
- Martín-Gómez, C., & Erduran, S. (2018). Understanding argumentation about socio-scientific issues on energy: a quantitative study with primary pre-service teachers in Spain. *Research in Science & Technological Education*, 36(4), 463-483.
- Mason, L., & Scirica, F. (2006). Prediction of students' argumentation skills about controversial topics by epistemological understanding. *Learning and instruction*, 16(5), 492-509.
- McCallin, R. C. (2015). Test administration. In *Handbook of test development* (pp. 583-600). Routledge.
- McConnell, M. M., St-Onge, C., & Young, M. E. (2014). The benefits of testing for learning on later performance. *Advances in Health Sciences Education*, 20(2), 305–320.
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276-282.
- McIntosh, M. J., & Morse, J. M. (2015). Situating and constructing diversity in semi-structured interviews. *Global qualitative nursing research*, 2, 2333393615597674.
- McNeill, K. L. (2011). Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year. *Journal of Research in Science Teaching*, 48(7), 793-823.
- McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *The journal of the Learning Sciences*, 15(2), 153-191.
- McNeill, K. L., & Krajcik, J. (2007). Middle school students’ use of appropriate and inappropriate evidence in writing scientific explanations. *Thinking with data*, 233-265.
- McPeck, J. (1981). *Critical thinking and education*. New York: St. Martin's.
- Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and instruction*, 14(2), 139-178.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741.
- Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent?. *Educational Measurement: issues and practice*, 8(1), 14-22.
- Meng, Y., Shi, Z., Zhang, X. (2022). Chuangxin kaocha xingshi peiyang gaojie siwei-2021 nian gaokao wuli Beijing juan di 15 ti de fenxi yu qishi [Innovating the form of examination and cultivating higher-order thinking—Analysis and enlightenment of the 15<sup>th</sup> item of the 2021 college entrance examination Physics Beijing volume]. *Shiyan jiaoxue yu yiqi*, (03), 12-13.
- Ministry of Education, P. R. China. (2017). *Physics curriculum standards for senior high school*. People’s Education Press.
- Ministry of Education, P. R. China. (2021). 2021 年全国高考报名人数 1078 万 [http://www.moe.gov.cn/jyb\\_xwfb/xw\\_zt/moe\\_357/2021/2021\\_zt12/meiti/202106/t20](http://www.moe.gov.cn/jyb_xwfb/xw_zt/moe_357/2021/2021_zt12/meiti/202106/t20)

- Ministry of Education, P. R. China. (2022). Science curriculum standards for compulsory education. People's Education Press.
- Mislevy, R. J. (2007). Validity by design. *Educational researcher*, 36(8), 463-469.
- Mohan, R. (2016). *Measurement, evaluation and assessment in education*. PHI Learning Pvt. Ltd..
- Morgan, D. L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of mixed methods research*, 1(1), 48-76.
- Morgan, D. L. (2014a). Pragmatism as a paradigm for social research. *Qualitative inquiry*, 20(8), 1045-1053.
- Morgan, D. (2014b). *Integrating qualitative and quantitative methods*. SAGE Publications, Inc.
- Mounce, H. (2002). *The two pragmatisms: from Peirce to Rorty*. Routledge.
- Münchow, H., Richter, T., von der Mühlen, S., & Schmid, S. (2019). The ability to evaluate arguments in scientific texts: Measurement, cognitive processes, nomological network, and relevance for academic success at the university. *British Journal of Educational Psychology*, 89(3), 501–523.
- Muthanna, A., & Sang, G. (2016). Undergraduate Chinese students' perspectives on Gaokao examination: Strengths, weaknesses, and implications. *International Journal of Research Studies in Education*, 5(2), 3-12.
- Nardi, P. M. (2014). *Doing survey research: a guide to quantitative methods* (3rd ed.).
- Nasir, N., Roseberry, A. S., Warren, B., & Lee, C. (2006). Learning as a cultural process: Achieving equity through diversity. In K. Sawyer (Ed.), *Handbook for the learning sciences* (pp. 489–504). Cambridge: Cambridge University Press
- Naumenko, O. (2014). Comparison of various polytomous item response theory modeling approaches for task based simulation cpa exam data. AICPA 2014 Summer Internship Project.
- Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy and Practice*, 14(2), 149–170.
- Newton, P. E. (2016). Macro-and micro-validation: Beyond the 'five sources' framework for classifying validation evidence and analysis. *Practical Assessment, Research, and Evaluation*, 21(1), 12.
- Newton, P. E. (2017). *An approach to understanding validation arguments*. Report for the Office of Qualifications and Examinations Regulation.
- NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press.
- Ng Yee Ping, D. (2019). Assessment of Argumentation in Chemistry: A Model for Designing Items. *Argumentation in Chemistry Education*, 106-141.
- Nielsen, J. A. (2013). Dialectical features of students' argumentation: A critical review of argumentation studies in science education. *Research in Science Education*, 43(1), 371-393.
- Noroozi, O., Weinberger, A., Biemans, H. J., Mulder, M., & Chizari, M. (2013). Facilitating argumentative knowledge construction through a transactive discussion script in CSCL. *Computers & Education*, 61, 59-76.
- Norton-Meier, L. (2008). Creating border convergence between science and language: A case for the science writing heuristic. In *Science inquiry, argument and language* (pp. 13-24). Brill Sense.
- Nussbaum, E. M., Kardash, C. M., & Graham, S. (Ed.). (2005). The Effects of Goal Instructions and Text on the Generation of Counterarguments During Writing. *Journal of Educational Psychology*, 97(2), 157–169.

- Nussbaum, E. M., Sinatra, G. M., & Poliquin, A. (2008). Role of epistemic beliefs and scientific argumentation in science learning. *International Journal of Science Education*, 30(15), 1977-1999.
- OECD (2020), *Benchmarking the Performance of China's Education System, PISA*, OECD Publishing, Paris.
- Olson, G. M., Duffy, S. A., & Mack, R. L. (2018). Thinking-out-loud as a method for studying real-time comprehension processes. In *New methods in reading comprehension research* (pp. 253-286). Routledge.
- Organisation for Economic Co-operation and Development (OECD). (2010). *Education at a glance 2010: OECD indicators*. Paris: OECD.
- Osborne, J. (2002). Science without literacy: A ship without a sail?. *Cambridge Journal of Education*, 32(2), 203-218.
- Osborne, J. (2012). The role of argument: Learning how to learn in school science. In *Second international handbook of science education* (pp. 933-949). Springer, Dordrecht.
- Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education*, 25(2), 177-196.
- Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, 53(6), 821-846.
- Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction?. *Science Education*, 95(4), 627-638.
- Osborne, J., & Patterson, A. (2012). Authors' response to "For whom is argument and explanation a necessary distinction? A response to Osborne and Patterson" by Berland and McNeill. *Science Education*, 96(5), 814-817.
- Paek, I., & Cole, K. (2019). *Using R for Item Response Theory Model Applications* (1st ed.). Routledge.
- Polesel, J., Rice, S., & Dulfer, N. (2014). The impact of high-stakes testing on curriculum and pedagogy: A teacher perspective from Australia. *Journal of education policy*, 29(5), 640-657.
- Popham, W. J. (1987). The merits of measurement-driven instruction. *The Phi Delta Kappan*, 68(9), 679-682.
- Qi, L. (2004). Has a high-stakes test produced the intended changes?. In *Washback in language testing* (pp. 193-212). Routledge.
- Qi, L., (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education*, 14(1), 51-74.
- Qu, S. Q., & Dumay, J. (2011). The qualitative research interview. *Qualitative research in accounting & management*.
- Quinlan, C. L. (2020). Analysis of preservice teachers' lesson plans to determine the extent of transfer of argumentation. *International Journal of Science Education*, 42(7), 1207-1223.
- Raïche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch measurement transactions*, 19(1), 1012.
- Rapanta, C., Garcia-Mila, M., & Gilabert, S. (2013). What is meant by argumentative competence? An integrative review of methods of analysis and assessment in education. *Review of Educational Research*, 83(4), 483-520.
- Rapanta, C., & Macagno, F. (2016). Argumentation methods in educational contexts: Introduction to the special issue. *International Journal of Educational Research*, 79, 142-149.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen:

- Danish Institute for Educational Research.
- Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., ... & Lewis, T. (2000). *A user's guide to MLwiN*. London: Institute of Education, 286.
- Ren, & Li. (2012). Research progress of Toulmin's model in science education. *Elementary & Secondary Schooling Abroad*, 9, 28-34.
- Reznitskaya, A., & Gregory, M. (2013). Student thought and classroom language: Examining the mechanisms of change in dialogic teaching. *Educational Psychologist*, 48(2), 114-133.
- Reznitskaya, A., Kuo, L. J., Clark, A. M., Miller, B., Jadallah, M., Anderson, R. C., & Nguyen-Jahiel, K. (2009). Collaborative reasoning: A dialogic approach to group discussions. *Cambridge journal of education*, 39(1), 29-48.
- Reith, M., & Nehring, A. (2020). Scientific reasoning and views on the nature of scientific inquiry: testing a new framework to understand and model epistemic cognition in science. *International Journal of Science Education*, 42(16), 2716-2741.
- Rivard, L. P., & Straw, S. B. (2000). The effect of talk and writing on learning science: An exploratory study. *Science education*, 84(5), 566-593.
- Rockstuhl, T., & Lievens, F. (2021). Prompt-specificity in scenario-based assessments: Associations with personality versus knowledge and effects on predictive validity. *Journal of Applied Psychology*, 106(1), 122.
- Romine, W. L., Sadler, T. D., & Kinslow, A. T. (2017). Assessment of scientific literacy: Development and validation of the Quantitative Assessment of Socio-Scientific Reasoning (QuASSR). *Journal of Research in Science Teaching*, 54(2), 274-295.
- Romine, W. L., Sadler, T. D., Dauer, J. M., & Kinslow, A. (2020). Measurement of socio-scientific reasoning (SSR) and exploration of SSR as a progression of competencies. *International Journal of Science Education*, 42(18), 2981-3002.
- Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten Benefits of Testing and Their Applications to Educational Practice. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 55). Elsevier Inc.
- Russell, T., & Aydeniz, M. (2013). Traversing the divide between high school science students and sophisticated nature of science understandings: A multi-pronged approach. *Journal of Science Education and Technology*, 22(4), 529-547.
- Ryu, S. (2011). *The appropriation of argumentation norms in a classroom community*. University of California, Los Angeles.
- Ryu, S., & Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education*, 96(3), 488-526.
- Sadler, T. D. (2006). Promoting discourse and argumentation in science teacher education. *Journal of Science Teacher Education*, 17(4), 323-346.
- Sadler, T. D., & Donnelly, L. A. (2006). Socioscientific argumentation: The effects of content knowledge and morality. *International Journal of Science Education*, 28(12), 1463-1488.
- Sampson, V., & Clark, D. (2006). Assessment of argument in science education: A critical review of the literature. <https://repository.isls.org/handle/1/3571>
- Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science education*, 92(3), 447-472.
- Sampson, V., Enderle, P., & Grooms, J. (2013). Argumentation in science education: Helping students understand the nature of scientific argumentation so they can meet the new science standards. *The Science Teacher*, 80(5), 30-33.
- Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *The journal of the learning sciences*, 12(1), 5-51.

- Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science education*, 89(4), 634-656.
- Sandoval, W. A., & Millwood, K. A. (2007). What can argumentation tell us about epistemology? In *Argumentation in science education* (pp. 71-88): Springer.
- Schwarz, B. B., Neuman, Y., Gil, J., & Ilya, M. (2003). Construction of collective and individual knowledge in argumentative activity. *Journal of the Learning Sciences*, 12(2), 219–256.
- Schwarz, B. B., & Baker, M. J. (2016). *Dialogue, argumentation and education: History, theory and practice*. Cambridge University Press.
- Scheerens, J., Glas, C. A., Thomas, S. M., & Thomas, S. (2003). *Educational evaluation, assessment, and monitoring: A systemic approach*. Taylor & Francis.
- Sengul, O., Enderle, P. J., & Schwartz, R. S. (2020). Science teachers' use of argumentation instructional model: linking PCK of argumentation, epistemological beliefs, and practice. *International Journal of Science Education*, 42(7), 1068-1086.
- Shao, Z., Pang, W. (2016). Gaokao chengji xingbie chayi yanjiu de huigu yu zhanwang [Retrospect and Future Research on Gender Differences in College Entrance Examination]. *Huadong shifan daxue xuebao (jiaoyu kexue ban)*, 34(1), 69.
- Shaw, V. F. (1996). The cognitive processes in informal reasoning. *Thinking & Reasoning*, 2(1), 51-80.
- Shaw, S., & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters, Special*, (3), 1-44.
- Shi, Y. (2019). Enhancing evidence-based argumentation in a Mainland China middle school. *Contemporary Educational Psychology*, 59, Article 101809.
- Shi, Y. (2020). Talk about evidence during argumentation. *Discourse Processes*, 57(9), 770-792.
- Shi, F., Wang, L., Liu, X., & Chiu, M. H. (2021). Development and validation of an observation protocol for measuring science teachers' modeling-based teaching performance. *Journal of Research in Science Teaching*.
- Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M. W. (2016). Advances in the science of assessment. *Educational Assessment*, 21(1), 34-59.
- Siler, S. A., & Klahr, D. (2016). Effects of terminological concreteness on middle-school students' learning of experimental design. *Journal of Educational Psychology*, 108(4), 547.
- Smith, C., Wiser, M., Anderson, C.W., & Krajcik, J. (2006). Implications for children's learning for assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement*, 14(1&2), 1–98.
- Song, & Wang. (2018). Scientific argumentation teaching: new progress internationally. *Comparative Education Review*, 7.
- Spencer-Rodgers, J., Williams, M. J., & Peng, K. (2010). Cultural differences in expectations of change and tolerance for contradiction: A decade of empirical research. *Personality and Social Psychology Review*, 14(3), 296–312.
- Stein, N. L., & Miller, C. A. (1993). The development of memory and reasoning skill in argumentative contexts: Evaluating, explaining, and generating evidence. *Advances in instructional psychology*, 4, 285-335.
- Sugirin. (1999). *Exploring the comprehension strategies of EFL readers: A multi-method study*. ERIC Clearinghouse.
- Sun, B. (2019). Dui chuzhong wuli ketang kexuetanjiu “fenxi, lunzheng yu pinggu” xianzhuang de fenxi [Analysis of the current situation of “analysis, argumentation and evaluation” of scientific inquiry in junior high school physics classroom]. *Wuli jiaoshi*, 1.

- Szu, E., & Osborne, J. (2012). Scientific reasoning and argumentation from a Bayesian perspective. In *Perspectives on scientific argumentation* (pp. 55-71). Springer, Dordrecht.
- Tang, K. S., & Moje, E. B. (2010). Relating multimodal representations to the literacies of science. *Research in Science Education*, 40(1), 81-85.
- Tang, Y., Hu, Y. (2013). Gaozhongsheng kexuexingqu, kexuejiazhiguan ji kexuesuyang xingbie chayi de shizhengyanjiu [An empirical study on gender differences in high school students' scientific interest, scientific values and scientific literacy]. *Zhongguorenmindaxue jiaoyu xuekan*, (2), 98-107.
- Tashakkori, A., & Teddlie, C. (2009). Integrating qualitative and quantitative approaches to research. *The SAGE handbook of applied social research methods*, 2, 283-317.
- Tashakkori, A., Johnson, R. B., & Teddlie, C. (2020). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Sage publications.
- Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools*, 13(1), 12-28.
- Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. *Journal of mixed methods research*, 1(1), 77-100.
- Thorndike, R. L., and Hagen, E. (1986), *Cognitive Abilities Test: Examiner's Manual Form 4*, Chicago, IL: Riverside.
- Tilly, C. (2006). *Why?: What happens when people give reasons...and why*. Princeton, NJ: Princeton University Press
- Toulmin, S. (2003). *The Uses of Argument* (2nd ed.). Cambridge: Cambridge University Press.
- Tseng, A. S., Bonilla, S., & MacPherson, A. (2021). Fighting “bad science” in the information age: The effects of an intervention to stimulate evaluation and critique of false scientific claims. *Journal of Research in Science Teaching*, 58(8), 1152-1178.
- Van Eemeren, F. H., Grootendorst, R., & Kruiger, T. (1984). *The Study of Argumentation*: New York: Irvington 0829015485/9780829015485.
- Van Eemeren, F. H., Grootendorst, R., Johnson, R. H., Plantin, C., & Willard, C. A. (2013). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge.
- Van Eemeren, F. H., Henkemans, A. F. S., & Grootendorst, R. (2002). *Argumentation: Analysis, evaluation, presentation*: Routledge.
- Vansteenkiste, M., Soenens, B., Verstuyf, J., & Lens, W. (2009). What is the usefulness of your schoolwork? The differential effects of intrinsic and extrinsic goal framing on optimal learning. *Theory and Research in Education*, 7(2), 155-163.
- Von Aufschnaiter, C., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 45(1), 101-131.
- von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E. M., & Berthold, K. (2016). Judging the plausibility of arguments in scientific texts: A student—Scientist comparison. *Thinking & Reasoning*, 22(2), 221–249.
- Vygotsky, L. S. (1962). *Thought and word*.
- Vygotsky, L. S. (1987). *The collected works of LS Vygotsky: the fundamentals of defectology* (Vol. 2): Springer Science & Business Media.
- Walton, D. N. (1989). Dialogue theory for critical thinking. *Argumentation*, 3, 169 – 184.
- Wang, J., & Buck, G. (2015). The relationship between Chinese students' subject matter knowledge and argumentation pedagogy. *International Journal of Science Education*, 37(2), 340-366.



- Wang, W. C., Cheng, Y. Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement*, 65(1), 5-27.
- Wang, J., Li, Q., & Luo, Y. (2020). Physics identity of Chinese students before and after gaokao: The effect of high-stake testing. *Research in Science Education*, 1-15.
- Wang, Y., Peng, X. (2012). Gaokao qikao beihou de chengxiang jiaoyu chaju fenxi jiqi duice yanjiu [Analysis of the gap between urban and rural education behind the abandonment of the college entrance examination and its countermeasures]. *Neimenggu shifan daxue xuebao: jiaoyu kexue ban*, (4), 95-97.
- Wang, Z., & Song, G. (2021). Towards an assessment of students' interdisciplinary competence in middle school science. *International Journal of Science Education*, 1-24.
- Weinert, F. E. (1999, April). Concepts of competence. OFS.
- Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Hogrefe & Huber Publishers.
- Wellington, J., & Osborne, J. (2001). *Language and literacy in science education*. McGraw-Hill Education (UK).
- West, A. (2010). High stakes testing, accountability, incentives and consequences in English schools. *Policy & politics*, 38(1), 23-39.
- Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge.
- Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 46(6), 716-730.
- Wilson, M. (2013). Using the concept of a measurement system to characterize measurement models used in psychometrics. *Measurement*, 46(9), 3766-3774.
- Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi delta kappan*, 75(3), 200-213.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press.
- Wright, B. D., & Mok, M. (2000). Understanding Rasch measurement: Rasch models overview. *Journal of applied measurement*.
- Wu, Y., Guo, Y. (2019). Zhongxuesheng kexue xuexi de xingbie chayi yu kecheng yingdui-jiyu PISA 2015 zhongguo si shengshi de shuju fenxi [Gender Differences in Middle School Students' Science Learning and Corresponding Strategies of Curriculum——Based on PISA 2015 Data Analysis of Four Provinces and Cities in China]. *Huadong shifan daxue xuebao (jiaoyu kexue ban)*, 37(5), 115.
- Wu, & Liu. (2017). The trend of scientific argumentation's assessment research from an international perspective. *Bulletin of Biology*(5), 4.
- Wyse, A. E. (2013). Construct maps as a foundation for standard setting. *Measurement: Interdisciplinary Research and Perspectives*, 11(4), 139-170.
- Xia, X., Jiang, Y., Bai, Y., Ye, X., Zheng, W. (2022). 2021 nian gaokao wuli shiti Zhong kexue tuili nengli kaocha de dingliang fenxi [Quantitative analysis of the assessment of scientific reasoning ability in 2021 college entrance examination Physics items]. *Wuli tongbao*, (04),134-136+152.
- Xie, Y., Hample, D., & Wang, X. (2015). A cross-cultural analysis of argument predispositions in China: Argumentativeness, verbal aggressiveness, argument frames, and personalization of conflict. *Argumentation*, 29(3), 265-284.
- Xin, T. (2016). Xuesheng fazhanhexinsuyang yanjiu ying zhuyi jige wenti [Several key points in the research of students' development of core literacy]. *Huadong shifan daxue xuebao (jiaoyu kexue ban)*, (1), 6-7.
- Xue, E, & Li, J. (2022). What is the value essence of “double reduction”(Shuang Jian) policy



- in China? A policy narrative perspective. *Educational Philosophy and Theory*, 1-10.
- Yaman, F. (2020). Pre-service science teachers' development and use of multiple levels of representation and written arguments in general chemistry laboratory courses. *Research in Science Education*, 50(6), 2331-2362.
- Yan, C. (2015). 'We can't change much unless the exams change': Teachers' dilemmas in the curriculum reform in China. *Improving Schools*, 18(1), 5-19.
- Yang, W. T., Lin, Y. R., She, H. C., & Huang, K. Y. (2015). The effects of prior-knowledge and online learning approaches on students' inquiry and argumentation abilities. *International Journal of Science Education*, 37(10), 1564-1589.
- Yao, S. Y., Wilson, M., Henderson, J. B., & Osborne, J. (2015). Investigating the Function of Content and Argumentation Items in a Science Test: A Multidimensional Approach. *Journal of applied measurement*, 16(2), 171-192.
- Ye, H. (2011). Gaokao de chengxiang chayi ji duice yanjiu [A study on the differences between urban and rural areas of the college entrance examination and its countermeasures]. *Zhongguo gaojiao yanjiu*, (4), 23-25.
- Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, 8(2), 125-145.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, 30(3), 187-213.
- Yore, L., Bisanz, G. L., & Hand, B. M. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. *International journal of science education*, 25(6), 689-725.
- Yu, L., & Suen, H. K. (2005). Historical and contemporary exam-driven education fever in China. *KEDI Journal of Educational Policy*, 2(1).
- Yu, S., Chen, B., Levesque-Bristol, C., & Vansteenkiste, M. (2018). Chinese education examined via the lens of self-determination. *Educational Psychology Review*, 30(1), 177-214.
- Yun, S. M., & Kim, H. B. (2015). Changes in students' participation and small group norms in scientific argumentation. *Research in Science Education*, 45(3), 465-484.
- Yvonne Feilzer, M. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of mixed methods research*, 4(1), 6-16.
- Zeidler, D. L., & Sadler, T. D. (2007). The role of moral reasoning in argumentation: Conscience, character, and care. In *Argumentation in science education* (pp. 201-216). Springer, Dordrecht.
- Zeidler, D. L., Herman, B. C., Ruzek, M., Linder, A., & Lin, S. S. (2013). Cross-cultural epistemological orientations to socioscientific issues. *Journal of Research in Science Teaching*, 50(3), 251-283.
- Zhang. (2018). Preliminary research of the assessment of middle school students' scientific argumentation ability. Northeast Normal University,
- Zhang, C., Chen, Y. (2018). Guanyu gaozhong wuli shishi "kexue lunzheng jiaoxue" de diaoyan yu sikao [Investigation and Reflection on the Implementation of "Scientific argumentation Teaching" in High School Physics]. *Wuli jiaoshi*, 39(5), 2-5.
- Zhang, D., Li, X., & Xue, J. (2015). Education inequality between rural and urban areas of the People's Republic of China, migrants' children education, and some implications. *Asian Development Review*, 32(1), 196-224.
- Zhang, H., Zhang, J. (2022). Gaokao wuli shiti de qingjinghua tezheng yanjiu [Research on the scenario characteristics of college entrance examination Physics questions]. *Wuli jiaoshi*, (03), 75-79.

- Zhao, X., Selman, R. L., & Haste, H. (2015). Academic stress in Chinese schools and a proposed preventive intervention program. *Cogent Education*, 2(1), 1-14.
- Zhao, G., Zhao, R., Li, X., Duan, Y., & Long, T. (2021). Are preservice science teachers (PSTs) prepared for teaching argumentation? Evidence from a university teacher preparation program in China. *Research in Science & Technological Education*, 1-20.
- Zheng, Zhang, & Zhang. (2019). High school students' scientific argumentation ability-assessment based on Rasch analysis. *Physics Teacher*, 40(1), 2.
- Zhou, H., Yu, G., Cai, Q., Dong, Y., Cao, W. (2022). Youhua qingjing sheji luoshi “siceng” “siyi”—2021 nian quanguo gaokao wuli Guangdong juan de Pingxi ji jiaoxue qishi [Optimizing scenario design and implementing "Four Layers" and "Four Wings"—— Analysis and teaching implications of the 2021 national college entrance examination Physics Guangdong volume]. *Wuli jiaoxue*, (02),67-71+11.
- Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 39(1), 35-62.
- Žukauskas, P., Vveinhardt, J., & Andriukaitienė, R. (2018). Philosophy and paradigm of scientific research. *Management culture and corporate social responsibility*, 121.

## Appendix 1 Comparison of Curriculum 2003 and 2017

2003 Curriculum	2017 Curriculum
<p><b>Curriculum overall aim:</b> Learn the basic knowledge and skills of physics necessary for life-long development, understand the application of these knowledge and skills in life and production, and pay attention to the current situation and development trend of science and technology.</p> <p>Learn scientific inquiry methods, develop independent learning ability, develop good thinking habits, and be able to use physical knowledge and scientific inquiry methods to solve some problems.</p> <p>Develop curiosity and thirst for knowledge, develop interest in scientific exploration, have a scientific attitude and scientific spirit of adhering to truth, be brave in innovation, and seek truth from facts, and have a sense of social responsibility to revitalize China and serve science to mankind.</p> <p>Understand the interaction between science and technology, economy and society, understand the relationship between man, nature and society, and have a sense of sustainable development and a global concept.</p> <p><b>Specific aim:</b> <b>(1) Knowledge and skills</b> 1. Learn the basic knowledge of physics, understand some basic concepts and laws of the structure, interaction and motion of matter, and understand the basic viewpoints and ideas of physics. 2. Understand the status and function of experiments in physics, master some basic skills of physical experiments, be able to use basic experimental instruments, and be able to complete some physical experiments independently. 3. Get a preliminary understanding of the development process of physics, pay attention to the main achievements and development trends of science and</p>	<p><b>Curriculum aim:</b> 1. Form material concepts, motion and interaction concepts, energy concepts, etc., which can be used to explain natural phenomena and solve practical problems.</p> <p>2. Have the awareness and ability to construct models; be able to use scientific thinking methods to conduct scientific reasoning, find out rules, and form conclusions on relevant issues from both qualitative and quantitative aspects; have the awareness of using scientific evidence and the ability to evaluate scientific evidence, Able to use evidence to describe, explain and predict research issues; have a critical thinking awareness, be able to question boldly based on evidence, think about problems from different perspectives, and pursue technological innovation.</p> <p>3. Have a sense of scientific inquiry, be able to discover problems and put forward reasonable conjectures and hypotheses in observations and experiments; have the ability to design inquiry plans and obtain evidence, correctly implement inquiry plans, use different methods and means to analyze and process information, describe and explain the results and trends of inquiry; have the willingness and ability to communicate, and be able to accurately express, evaluate and reflect on the process and results of inquiry.</p> <p>4. Be able to correctly understand the nature of science; have the curiosity and thirst for knowledge in learning and researching physics, be able to actively cooperate with others, respect others, be able to express their own opinions based on evidence and logic, seek truth from facts, not superstitious authority; care about domestic and foreign science and technology Development status and trends, understanding of physical research and application of physical results should follow ethical norms, understand the</p>

technology, and the impact of physics on economic and social development.

4. Pay attention to the connection between physics and other disciplines, know some application areas related to physics, and be able to try to use relevant physical knowledge and skills to explain some natural phenomena and problems in life.

**(2) Process and method**

1. Experience the process of scientific inquiry, understand the meaning of scientific inquiry, and try to apply scientific inquiry methods to study physical problems and verify physical laws.
2. Through the learning process of physical concepts and laws, understand the research methods of physics, and recognize the role of physical experiments, physical models and mathematical tools in the development of physics.
3. Able to plan and regulate their own learning process, solve some physical problems encountered in learning through their own efforts, and have a certain ability to learn independently.
4. Participate in some scientific practice activities, try to express your own opinions after thinking, and try to solve some practical problems related to production and life by using physical principles and research methods.
5. Have certain questioning ability, information collection and processing ability, analysis, problem-solving ability and communication and cooperation ability.

**(3) Emotional Attitudes and Values**

1. Can appreciate the wonder and harmony of nature, develop curiosity and thirst for knowledge in science, be willing to explore the mysteries of nature, and experience the hardships and joys of exploring the laws of nature.
2. Have the enthusiasm to participate in scientific and technological activities, have the awareness of applying physical knowledge to life and production practice, and have the courage to explore physical problems related to daily life.

relationship between science, technology, society and environment, and have a sense of responsibility to protect the environment, save resources, and promote sustainable development.

<p>3. Have the scientific attitude and scientific spirit of daring to adhere to the truth, be brave in innovation and seeking truth from facts, and have the consciousness of judging whether the relevant information of the mass media is scientific or not.</p> <p>4. Have the spirit of taking the initiative to cooperate with others, have the desire to communicate your own opinions with others, have the courage to insist on correct viewpoints, have the courage to correct mistakes, and have team spirit.</p> <p>5. Understand and appreciate the contribution of physics to economic and social development, pay attention to and think about hot issues related to physics, have the awareness of sustainable development, and be able to contribute to the sustainable development of society within the scope of ability.</p> <p>6. Care about the current situation and trends of domestic and foreign scientific and technological development, have a sense of mission and responsibility to revitalize China, and have a sense of serving science to mankind.</p>	
<p><b>Content standard (Scientific inquiry and physical experiment ability requirements):</b></p> <p><b>Find a problem</b> Can discover problems related to physics. Formulate these questions more clearly from the point of view of physics. Recognize the significance of identifying and asking questions.</p> <p><b>Assumptions and hypothesis</b> Make assumptions about how problems are solved and answers to questions. Predict the results of physical experiments. Recognize the importance of assumptions and hypothesis.</p> <p><b>Develop plans and design experiments</b> Know the purpose of the experiment and the existing conditions and formulate the experimental plan. Try to choose the experimental method and the required equipment. Consider the variables of the experiment and how to control them. Recognize the role of planning.</p>	<p><b>Key competences:</b></p> <p><b>Physical concept</b> The concept of physics is the basic understanding of matter, motion and interaction, energy, etc. formed from the perspective of physics; it is the refinement and sublimation of physical concepts and laws in the mind; it is from the perspective of physics to explain natural phenomena and solve practical problems. Base. The concept of physics mainly includes the concept of matter, the concept of motion and interaction, the concept of energy and other elements.</p> <p><b>Scientific thinking</b> Scientific thinking is a way of understanding the essential properties, internal laws and interrelationships of objective things from the perspective of physics; it is an abstract and general process of constructing physical models based on empirical facts; it is the specific application of methods such as analysis and synthesis,</p>

<p><b>Conduct experiments and gather evidence</b>  Collect data in multiple ways.  Carry out the experimental operation according to the instructions and can use the basic experimental equipment.  Record experimental data truthfully and know the significance of repeatedly collecting experimental data.  Awareness of safe operation.  Recognize the importance of scientifically collecting experimental data.</p> <p><b>Analysis and Argumentation</b>  Analyze and process experimental data.  Attempt to draw conclusions based on experimental phenomena and data.  Interpret and describe experimental results.  It is important to recognize the importance of conducting analytical arguments in experiments.</p> <p><b>Evaluate</b>  Attempt to analyze the difference between hypothesis and experimental results.  Pay attention to unresolved contradictions in inquiry activities and discover new problems.  Lessons learned to improve inquiry programmes.  Recognize the significance of assessment.</p> <p><b>Exchange and cooperation</b>  Able to write experimental research reports.  Pay attention to both adherence to principles and respect for others in cooperation.  Have a spirit of cooperation.  Recognize the importance of communication and cooperation.</p>	<p>reasoning and argumentation in the field of science; It is the ability and character to question and criticize, test and revise different views and conclusions based on factual evidence and scientific reasoning, and then put forward creative opinions. Scientific thinking mainly includes the elements of model construction, scientific reasoning, scientific argumentation, questioning and innovation.</p> <p><b>Scientific inquiry</b>  Scientific inquiry refers to asking physical questions based on observations and experiments, forming conjectures and hypotheses, designing experiments and formulating plans, acquiring and processing information, drawing conclusions and explanations based on evidence, and communicating, evaluating, and reflecting on the process and results of scientific inquiry Ability.  Scientific inquiry mainly includes questions, evidence, explanation, communication and other elements.</p> <p><b>Scientific attitude and responsibility</b>  Scientific attitude and responsibility refer to the inner drive to explore nature gradually formed on the basis of understanding the nature of science and the relationship between science, technology, society and the environment, a rigorous, serious, realistic and persevering scientific attitude, as well as abiding by ethical norms and protecting the environment And promote a sense of responsibility for sustainable development.  Scientific attitude and responsibility mainly include the elements of scientific nature, scientific attitude and social responsibility.</p>
---	--

## Appendix 2 Ferrara and Lai's (2015) validation framework

Table 31.1 Examples of Claims and Evidence in Each Testing Program Step and Responsibilities for Contributing to an IJA Report

<i>Testing program step</i>	<i>Responsibility for completing the step</i>	<i>Example claim</i>	<i>Documentation of claims and evidence</i>	<i>Expertise and responsibility for documentation</i>	<i>Primary audience for the documentation</i>
Determination of testing program policies and articulation of intended interpretations and uses of test scores	Sponsors set policy and intentions, program managers and staff articulate the policy and intended interpretations and uses of test scores	Examinees who reach a cut score are proficient in the tested content or ready for college and work; candidates are adequately prepared to work successfully in a professional area	Before test implementation: Public statements and documents After implementation: All documentation	Expertise (i.e., authority): Program sponsors Responsibility: Program managers and staff	Test users Test takers
Test design and development	Program managers and staff Testing program contractors	Item development procedures produce items that elicit evidence of targeted content knowledge and skills	For example, documents that describe item developer training, procedures and criteria; research evidence that the items elicit targeted knowledge and skills	Expertise and responsibility: Program managers and staff Testing program contractors	Program managers and staff
Test implementation	Test administrators	The test administration followed prescribed procedures: It was orderly and free of distractions and no security breaches or cheating occurred	For example, test administration manuals with prescribed procedures; no reports of upset to the administration environment and procedures; no evidence of security breaches or cheating	Expertise: Test administrators Responsibility: Program managers and staff	Program managers and staff Test users
Response scoring (i.e., professional and machine scoring)	Program managers and staff Testing program contractors	Scoring processes are consistent and accurate for all examinees, minimizing errors and potential bias	For example, documents that indicate adequate rater agreement and accuracy (e.g., scorer performance reports)	Expertise and responsibility: Program managers and staff Testing program contractors Specifically, scoring experts and scoring engine developers	Program managers and staff Test users

(Continued)

Table 31.1 (Continued)

<i>Testing program step</i>	<i>Responsibility for completing the step</i>	<i>Example claim</i>	<i>Documentation of claims and evidence</i>	<i>Expertise and responsibility for documentation</i>	<i>Primary audience for the documentation</i>
Technical analyses (e.g., item and test analysis, scaling, equating)	Program managers and staff Testing program contractors	Test scores are adequately and similarly reliable and generalizable for all examinee subgroups	Technical report: for example, reliability estimates and generalizability coefficients indicate score reliability and generalizability to support intended interpretations for all examinees	Expertise and responsibility: Program managers and staff Testing program contractors Specifically, psychometricians	Program managers and staff Test users
Delivery of scores and other feedback to examinees, candidates and other test users; training to support valid interpretations and uses of scores	Program managers and staff Testing program contractors	Scores and other feedback enable intended interpretations and uses; training is effective in enabling valid interpretations and uses	For example, score reports and interpretation guides, training materials For example, results of usability studies of the score reports	Expertise and responsibility: Program managers and staff Testing program contractors	Test users Test takers
Interpretations of score reports to guide decisions and take other actions	Test users	Test users interpret and use results in intended ways	For example, results from studies on test score use and associated positive and negative consequences	Test users	Program managers and staff Test takers



Table 31.2 Representation of Supporting and Disconfirming Evidence and Final Conclusion for One Claim: Item Development Procedures Produce Items That Elicit Evidence of Targeted KSAs

<i>Claim and dependent subclaims</i>	<i>Supporting evidence 1</i>	<i>Supporting evidence 2</i>	<i>Disconfirming evidence</i>	<i>Status of claim</i>
<b>Claim 1: Item development procedures produce items that elicit evidence of targeted knowledge, skills and abilities (KSAs)</b>				<b>Rejected</b>
Subclaim A: Item writers understand the assessment targets and how to write items to elicit the targeted KSAs	Item writer educational backgrounds and item writing experiences are extensive	Training materials indicate strong focus on KSAs; training evaluation responses indicate strong understanding of the KSAs and how items can target KSAs	—	Strongly supported
Subclaim B: Item development procedures and tools support item writers in focusing on the targeted KSAs and avoiding sources of construct-irrelevant variance	Item development guidelines and specifications include empirically supported information to align item features and response demands and targeted KSAs	Item templates specify target KSAs and allowable item features and response demands	—	Strongly supported
Subclaim C: Items are well aligned with development guidelines, specifications and templates	Independent alignment reviews indicate that items are well aligned to targeted content standards	In alignment reviews, multiple independent judgments agree about the levels 1, 2 and 3 depth-of-knowledge judgments for most items	Judgments indicate that some items intended to target high cognitive demands are aligned with lower cognitive demands	Weakly supported
Subclaim D: Items elicit targeted KSAs	Lower proficiency examinees generally do not respond successfully to higher-difficulty items; most item-total correlations are high (i.e., 0.30 and higher)	—	Think-aloud data indicate examinee response strategies that circumvent targeted KSAs	Rejected

Note: Status of claims can be evaluated as strongly supported, weakly supported, rejected or unexamined.

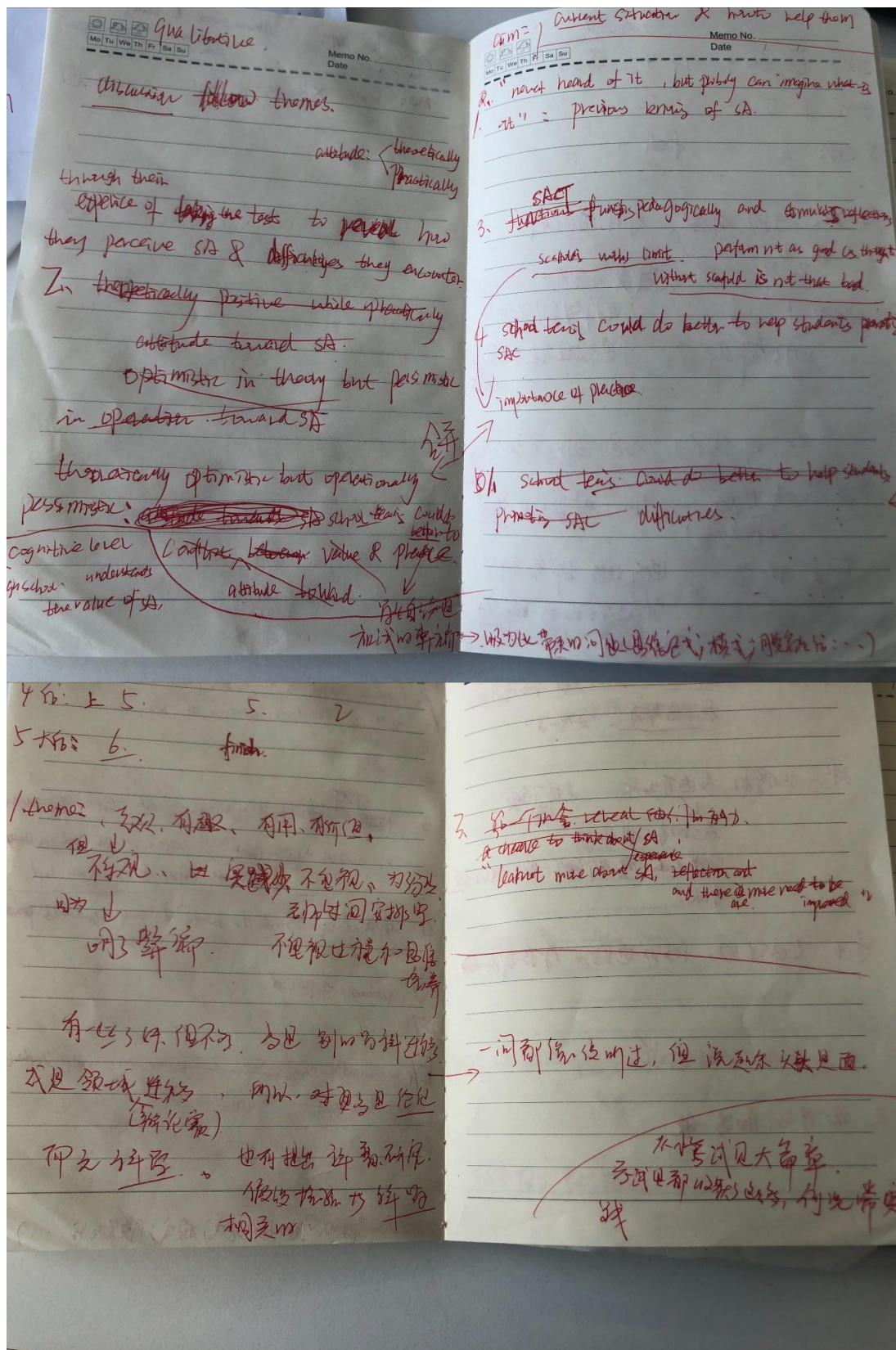
### **Appendix 3 Semi-structured follow-up interview outline**

- Did you know anything about SA? What did you know about it?
- How did you feel about taking the test?
- Are there any difference of your feeling between this test and other tests you usually take in school?
- What do you think the test was assessing?
- What characteristics do high quality SA possess? Why?
- Do you like argue or discuss with other people in your daily life? Why?
- Are there items that you find it's not clear or hard to understand?
- Which item(s) do you think are simpler than others? Why?
- Which item(s) do you think are more difficult? Why?
- Why you respond to this item this way? How did you think at that time? (wrong/poor answers or right answers on complicated item)
- Are you satisfied with your responses on this test? Why?
- Which item(s) do you think you performed well on? Why?
- Which item(s) do you think you performed not goods well on? Why?
- Under what kind of situations, or what changes do you expect you'd need so you can perform better than this time?
- Did you have any different understandings about SA after taking the assessment?
- Did you find the scaffold helpful?
- Could you please share your ideas about what if integrating SA into your school science learning?
- Do you have any concerns about yourself engaging in SA?
- Do you have any suggestions on how to better design the assessment?

## Appendix 4 Nvivo coding screenshot of think aloud data

2think aloud second pilot				
Name	Files	References	Created on	
<input type="radio"/> ignore evidence	1	1	2020/7/17 16:10	
<input type="radio"/> inconsistency	3	4	2020/7/14 15:40	
<input type="radio"/> influenced by dialogue	1	1	2020/7/17 9:05	
<input type="radio"/> misunderstand the test aim	2	3	2020/7/14 9:27	
<input type="radio"/> not confident in proposing claims	1	2	2020/7/16 10:05	
<input type="radio"/> role of scaffold	2	3	2020/7/14 14:59	
<input type="radio"/> task 1	0	0	2020/7/13 17:40	
<input type="radio"/> do not know how to respond to production items	1	1	2020/7/17 15:53	
<input type="radio"/> inconsistency	2	5	2020/7/13 18:29	
<input type="radio"/> information is not sufficient	1	1	2020/7/14 16:21	
<input type="radio"/> information too complex	4	5	2020/7/13 18:20	
<input type="radio"/> item stem is too long	1	1	2020/7/15 11:45	
<input type="radio"/> misunderstand item stem	4	4	2020/7/14 16:42	
<input type="radio"/> misunderstand the topic being argued	1	3	2020/7/13 18:09	
<input type="radio"/> reason is vague	1	1	2020/7/14 16:23	
<input type="radio"/> rebuttal is too simple	1	1	2020/7/14 16:26	
<input type="radio"/> talked better than wrote	1	1	2020/7/17 18:11	
<input type="radio"/> unclear item stem	1	2	2020/7/13 18:34	
<input type="radio"/> without paying attention to the problem being argued	2	2	2020/7/14 16:30	
<input type="radio"/> task 2	0	0	2020/7/13 17:40	
<input type="radio"/> confused about the topic being argued	1	1	2020/7/14 8:18	
<input type="radio"/> misunderstand item stem	3	4	2020/7/14 8:22	
<input type="radio"/> misunderstand the content knowledge	2	4	2020/7/14 8:15	
<input type="radio"/> unclear item options	1	1	2020/7/15 9:42	

Appendix 5 Theme construction drafts





有趣 - 接近生活、在真正思考、意识到了其价值  
新奇

1. 现有考试无的，一关注结果、高分与思维能力不挂钩、  
影响思维定式、不注意过程。但积极想赢得高<sup>分</sup>考  
学生。  
缺点

2. 之前有经历(生活中)、(测试中)得到一些

经历与的表现与理科课、偏向认知、何不  
是社会，更缺乏社会中的协作与科学本质  
生活中讨论、争辩、有一定即理科(字面意)  
给予一定帮助。  
只是不熟，但有这个能力。以制的成绩与认知 TA.

3. 4 个例子、讨论过程、引发讨论 -> 关于教育的例子、benefit

语言的逻辑性、提问题、做例子、So what (assessment)  
关注了哪些、  
关注了哪些、  
assessment

理由

Talmon's way of thinking

inconsistency

知识本身、Context  
knowledge, practice

4. 相信所见、信口直说、自由观点的碰撞

理由地推、对别人的观点、  
理解与操作、  
知识、  
得到与实际操作、  
范围、  
可及

knowledge not difficult: 3, 4, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18

Theme 1: students' attitude toward SA	Codes: Interested in doing SACT 4, 5, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18 Items are close to daily life 2, 3, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 Feeling unfamiliar with SA and SACT 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 SACT assesses more on thinking ability compared to their normal test 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18 SA is valuable in school study and life 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 Implementation of SA in school can be difficult 3, 6, 8, 10
Theme 2: students' previous knowledge of SA	Code: Have a sense of the activity while unfamiliar with terms 3, 4, 5, 6, 10, 11, 12 Cross disciplinary experience 2, 3, 4, 6, 8, 9, 10, 12, 14, 15, 16 Have certain understanding on SA while inaccurate
Theme 3: what SACT brings to students	Sub-theme 3.1 Pedagogical function (without context) 3, 4, 7, 11, 14, 16 Codes: (without semantic map) 2, 3, 5, 7, 8 (importance of practice) 4, 6 The four elements of SA and their meaning 2, 6, 8, 9, 10, 11, 14, 16, 17, 18, 19, 20 Learn more about how to argue 3, 5, 8, 9, 10, 11, 14, 15, 16, 17, 18 Sub-theme 3.2 (high quality) 2, 3, 5, 8, 9, 12 Evoke reflection Codes: discussion: 7, 11, 12, 13, 17 Aware of the importance of improving their language skills 2, 3, 4, 7, 9, 10, 11, 13, 14, 16, 18 Realize their ignorance of 'Reason' 1, 2, 4, 5, 6, 8, 9, 10, 11, 16, 17, 18 Aware of the importance of proposing their own claim 3, 6, 11, 14, 18 Aware of the role of listening to others 2, 10, 11, 14, 17, 18
Theme 4: how school teaching influences their engagement in SA	Sub-theme 4.1 thinking habit Codes: Care more about the final answer rather than the thinking process 2, 9, 10, 18, 14, 13 "fear" of words 1, 2, 10, 11, 12, 13 Apply tips to select Multiple Choice items rather than think over 5, 6, 12, 12 Sub-theme 4.2 little opportunity provided for engaging in SA Codes: assessment no SA now 3, 8, 9, 10, 12, 14, 18, 19 The aim of school teaching is high score on examinations 1, 10, 12, 18, 14, 14 Teaching schedule is too tight for the activity 2, 9, 12
Theme 5: what affect students' performance on SA	Sub-theme 1.1 common problems in students' SA Codes: Believe in what is provided 5, 8, 10, 16 (believe in what is provided) Rely on intuition 4, 5, 6, 8, 10, 11, 13, 17, 16 Separate formulas from physics problem understanding 1, 3, 4, 5, 8, 10, 11, 9, 10, 17, 12 Care more about the correct answer 9, 10, 13 Easily be shaken by other's claim 1, 2, 3, 10, 11 Sub-theme 1.2 difficulties of engaging in SA Codes: 3, 4, 12, 14, 14, 9, 10, 13, more 10 Confound reason and evidence; Weak understanding of reason Difficulty in rebutting others; Difficulty in writing Not confident about their verbal skill 2, 3, 4, 7, 11, 14, 16, 18 Awareness of the activity 5 Insufficient SA knowledge 2, 3, 4, 7, 8, 9, 11 Insufficient content knowledge 2, 3, 4, 7, 8, 11

fail to find < Psychol: 2, 3, 5, 6, 7, 8, 10, 12  
 hard to say/write: 3, 4, 9, 11, 12, 14, 18  
 maintain: 1, 10, 12, 5, 2, 3  
 13, 14, 17, 18  
 19, 20

## Appendix 6 Nvivo coding screenshot of follow-up interview data

main data Search Project

Name	Files	References
<input type="radio"/> content knowledge	1	1
<input type="radio"/> physchol influence	3	4
<input type="radio"/> the subject they choose influence confidence on verbal skill	2	4
<input type="radio"/> intuition	8	12
<input type="radio"/> know semantic meaning but cannot apply	7	10
<input type="radio"/> logical thinking	5	6
<input type="radio"/> MC question	7	10
<input type="radio"/> new understanding about the process of SA	10	15
<input type="radio"/> not confident because of school score	1	1
<input type="radio"/> open-ended question	5	6
<input type="radio"/> oral or written skill awareness	10	14
<input type="radio"/> personality	8	12
<input type="radio"/> previous sense of SA or related	11	24
<input type="radio"/> reason awareness	8	11
<input type="radio"/> rebuttal is difficult	8	11
<input type="radio"/> SA absent from current school learning	5	5
<input type="radio"/> SA is valuable for study	10	15
<input type="radio"/> scaffold helps	7	10
<input type="radio"/> scenario familiarity	11	11
<input type="radio"/> scenario misunderstanding	2	2
<input type="radio"/> school grade	8	12
<input type="radio"/> seperate calculation from physics learning	5	5
<input type="radio"/> suggestion	8	10

In Codes Code to physchol influence (Codes\main data\influencing factors)

## Appendix 7 Codes and themes of thematic analysis

### Theme 1

Code	Illustrative quotes
Felt it interesting to be engaged in SA	<p><i>“...the first reaction when I saw the test paper was excited...I am tired of the items we did in school test, and I felt that questions on this test was what real Physics should be like in my mind.... these items are special and apparently thinking ability is needed to analyze them...” (M8)</i></p> <p><i>“...these are not accessible in our usual study, and I felt it’s quite novel...we didn’t have this kind of discussion in school...I felt excited when got this test, it’s like finally I got a chance to present my, uh, to try these stuff...” (F14)</i></p>
Had more interest in thinking about close to life questions	<p><i>“...most of these scenarios are quite close to life, such as...sometimes I also thought about these questions when I saw these phenomena, but seldom paid attention to it...it’s quite interesting to think over these questions...” (F12)</i></p> <p><i>“I really think we come to school is to solve more (problems in real life), to know more scientific knowledge that happens in our life...for example, if something broken in my house, like electrical stuff, I can help my parents to fix it rather than to repeat the examination items once and once again on the text book, some of them, you know, too far away from our life and it’s just hard to imagine them” (M1)</i></p> <p><i>“It is quite interesting to think over these little problems in real life, although I cannot really figure out why for some of them, but you know, just interesting” (M2)</i></p>
SACT assessed more on thinking process compared with their normal test	<p><i>“Our normal examination is more about calculation and using formulas to answer a question, and more about the knowledge on the textbook. But this test gave me the feeling of assessing my understanding, and the ability to analyze a phenomenon or problem. We never met this kind of test before, and its more about thinking process” (M3)</i></p> <p><i>“...unlike our normal examinations that assesses knowledge and ask us to figure out the</i></p>



	<p><i>final answer follow the certain way we have been taught, this test is closer to real life and more practical, and cares more about the logic and thinking ability...like task 4, it provides several information and we need to propose and justify our claim based on these information, which is quite flexible and we have to pay attention to our own thinking...what is assessed here is like a whole set of stuff that is quite systematic” (F3)</i></p> <p><i>“The test we usually do is like just throw out the answer and it doesn’t emphasis on the process of your argument, but these items need you to show your thinking to others. I feel there is no standard right answers to these questions, it’s like you just need to justify your idea and to make it coherent. But our normal examination needs us to give right answer, it is not acceptable to just justify yourself” (F13)</i></p>
<p>Felt SA is valuable in school study and life</p>	<p><i>“These items are based on the content knowledge that we have learnt at school, so they kind of offered a different way of assessing or gaining new understanding about what we have learnt through asking us to justify our claim toward a problem” (F4)</i></p> <p><i>“In study, it is necessary to have the thinking of SA, no matter what you learn, if you want to really understand it and apply it, you have to have this kind of thinking ability...and another reason for why I like it is that I want to enhance or improve the reasonability of my own argument by arguing with others” (M8)</i></p> <p><i>“...it is important not only in study but also in life, if one possessed this ability, he or she can make right and reasonable decision or judgement when encountering various problems...” (M5)</i></p> <p><i>“...it is helpful for daily life as well, since we always need to have our own claim when talking with others...I feel it will be useful in our near future...after we entered into university, it will be useful for we cannot always use what is on the textbook...” (M11)</i></p>
<p>Never experienced SA in current achievement assessment</p>	<p><i>“...seldom have this kind of activity...probably because it is not included in the school examination and college entrance examination,</i></p>

	<p><i>but I feel it will be useful after high school...” (M11)</i></p> <p><i>“Actually, the examination items we usually take, and what we learned on the classroom, it’s like using formulas to get the answer... In most cases, these learning (take examinations) experience is like enabling us with muscle memory and the ability to control symbols, they do not make connection with real life or with logic” (M8)</i></p> <p><i>“Our examination doesn’t assess this kind of ability and we have the feeling that they care more about the final answer that is derived through applying a combination of formulas” (F11)</i></p>
Cross area experience	<p><i>“Teachers never taught us, but I watched some TV shows, like the debate competition, it is quite like that. I also like to watch TV shows about popular science and those scientists are amazing and I felt that it is like an activity between (them)” (M3)</i></p> <p><i>“Never heard of the term, but I think I know something about it. I used to like reading books related to science and learnt their way of telling a story by present their argument” (F13)</i></p>
Have a sense of SA while unfamiliar with the term	<p><i>“No, I never heard of it. But I after I did the test, I feel like it’s like the debate competition held in our school...and I kind of went through similar process when answering Political questions, but I didn’t know that is scientific argumentation until I did the test. It has something to do with way of thinking anyway” (M9)</i></p> <p><i>“We never learnt it before, but I feel like it has something to do with logical thinking, and it’s like propose some hypothesis to a phenomenon and explain that phenomenon” (M6)</i></p> <p><i>“I know nothing about SA, it’s like writing an article in Chinese, such as the claim, evidence</i></p>

*etc., and my Chinese is not good, so you know, I know very few about argumentation” (M1)*

*Theme 2*

<b>Code</b>	<b>Illustrative quotes</b>
Aware of the four elements and their meaning	<p><i>“...previously I didn’t have a clear understanding toward these four (elements), but now I think I understand them and the logic between them” (F10)</i></p>
Learned more about the process of argumentation	<p><i>“I didn’t know these terms and their meaning before, especially in the school learning context. But now I think I not only understand these elements but have a clear understanding about the whole thing by finish the test. When I progressed to the next few items, I would check the first few items using what I learned from the next few items” (F2)</i></p> <p><i>“I felt it became more comfortable as I progressed with more items because I got a sense of how the whole system is working, like the logical relationship between these elements. So, I found myself have a deeper understanding toward SA after completing the test” (F16)</i></p>
Sparked discussion among students	<p><i>“...we discussed a lot with our classmates after class, and you know, the situation was quite interesting because people have their own idea and they all wanted to prove themselves to be right...” (F8)</i></p> <p><i>“I remember the next lesson is a seminar, and several of us didn’t listen to the seminar</i></p>

	<p><i>but we were discussing about the items. We were so happy when we were discussing since everyone shared their thoughts which is different and several students from other classes seated near us also joined our discussion...After this experience, I found that it is interesting when we discuss a problem with uncertain answers and close to life, and we also found it is important to think about the problem from other's side and try to understand what they meant"</i></p> <p><i>(F16)</i></p>
<p>Aware of the importance of improving their language skill</p>	<p><i>"I found it very important to better express myself to make it clear and understandable..." (F18)</i></p> <p><i>"...this activity does require people to organize their language in a good way..." (F15)</i></p>
<p>Realized their ignorance of "Reason"</p>	<p><i>"About the "reason", I never thought that it could be in this way. I mean, "reason" is like used a lot in our life to explain something, but I never thought it is used to better connect between claim and evidence. This also reflects that I am not rigorous enough in this aspect (SA)" (F14)</i></p> <p><i>"I didn't realize that "reason" can be used in this way and have this meaning, and it is a bit hard to understand at the beginning, but after I finish the test, I think now I understand it" (F10)</i></p>
<p>Aware of the importance of proposing their</p>	<p><i>"I realized that everyone could have different opinions on a problem, and one</i></p>

own claim	<i>should propose their argument that is closely related to the claim” (F18)</i>
Aware of the role of listening to others in SA	<p><i>“I was thinking that in an argumentation, we only need to prove ourselves is right and to emphasize our own claim. But now I realized that we should also think from other’s stand and to try to prove why other’s claim is wrong” (F4)</i></p> <p><i>“Sometimes we need to try to understand others’ opinion to better communicate with others on a problem, for it is always the case that what others said might not be all wrong although it is different from my opinion” (F16)</i></p>
Inconsistency	<p><i>“Now I have a very clear understanding toward reason and evidence, reason is used to explain why the evidence supports the claim...according to fact b and c, well, he said that “he felt the black ball falls faster”, you know, he used the word felt, which is quite subjective. This should not be an objective evidence that can justify himself....” (F3)</i></p> <p><i>“Well, I think his evidence is relevant to his claim. His evidence is “the mass of black ball is larger”, how to say it, just the evidence is relevant because he used a fact that is a characteristic the ball possesses to prove his claim. In terms of whether this evidence is right or not to support the claim, probably because the reason is not sufficient</i></p>

	<p><i>I think “facts b and c” are not evidences, because these two facts cannot get the claim.</i></p> <p><i>Evidence exists objectively, and it is real and don’t need extra decoration or explanation. But reason is what helps to construct an argument by observing or analyzing the evidence. ...his reason is fact a.” (M5)</i></p>
--	---

*Theme 3*

<b>Code</b>	<b>Illustrative quotes</b>
Believing in what was provided	<p><i>“I also thought that he cannot adjust the angle and cannot make it fall at a constant speed. But when I looked at the provided dialogue, he said that “adjust the angle to make it fall at a constant speed”, which is a given information, so it must can fall in a constant speed....” (F10)</i></p> <p><i>“I am not sure how he gets the conclusion, but there must a reason. Let me see, well, it must can hit the ball or he won’t say that.” (M10)</i></p> <p><i>“I had no idea why this information was provided, and I didn’t use it for I didn’t know how to use it...but I think all the information provided in the task must be useful and were supposed to be used...” (F1)</i></p>
Relying on intuition	<p><i>“I choose C because I felt A and C must be wrong. But it’s hard to explain, just rely on feeling” (F17)</i></p> <p><i>“...it’s like his evidence is not enough...it’s a feeling” (F7)</i></p> <p><i>“Sometimes I kind of know how to deal with it intuitively but cannot explain why” (M3)</i></p>
Separating formulas from physics problem understanding	<p><i>“That is not about thinking over a physics problem, but focuses on the usage of formulas and providing the final right answer...” (M11)</i></p> <p><i>“I found that our normal test usually assesses</i></p>

	<p><i>calculation using formulas and content knowledge, this one is more about understanding” (M3)</i></p> <p><i>“I don’t like remember formulas, and I don’t know which one to use when doing our normal test items...” (M10)</i></p> <p><i>“I really don’t want to do Physics items (calculation using formulas) so I ignored the provided information which is about formulas...I gave up this one, because calculation is required for this item....” (F5)</i></p>
Expressing their own idea was out of their comfort zone	<p><i>“I was always doubting about the claim I proposed in the process of argumentation...” (M9)</i></p> <p><i>“I don’t really know what claim I should propose, it’s just like both arguments provided in the task seem reasonable....” (F17)</i></p> <p><i>“It is not easy to have our own claim, especially when others’ claim seems plausible...I was always feeling the other side’s claim being right when I was supporting my own claim...” (M10)</i></p>
Confound evidence and reason	<p><i>“I am confused about the difference between reason and evidence, because you know, both is to explain the claim...” (F17)</i></p>
Difficulty in providing reason	<p><i>“I felt that evidence and reason is so similar, and I tend to provide the same answer to items that ask me ‘what is your evidence?’ and items that ask me ‘why this evidence support your claim’” (M12)</i></p> <p><i>“I think that reason should explain how the evidence is used to support the claim, but I found that it has been hard to connect between claim and evidence accurately. So, I felt that my argument is not rigorous enough” (F14)</i></p> <p><i>“I felt that the reason I provided is not enough, but I don’t really know how to construct a coherent reason to explain the relationship between claim and evidence” (F9)</i></p>
Difficulty in rebutting others	<p><i>“I have no idea about how to rebut others, the rebuttal I wrote is like reason...” (M7)</i></p> <p><i>“I felt that the Production of Rebuttal items</i></p>

	<i>were difficult...Previously, I thought I just need to know my own claim and to prove it is correct, but now, I need to think from another side and to prove why others are wrong. I don't know how to do that..." (F4)</i>
Felt it not easy to express it out clearly	<i>"I just don't know how to express it...I think I know, but I cannot tell..." (F18)</i>  <i>"...for some items, I just don't know how to say it and how to make it clearly articulated" (M11)</i>  <i>"...you know, my language skill is poor, so just hard to say it out" (M4)</i>
Insufficient content knowledge	<i>"...my Physics grade is not good, and I probably will do better if know more about the knowledge" (F5)</i>
Insufficient SA knowledge	<i>"it's because I am not familiar with SA, if we are taught about it, I think I can do better" (F16)</i>
Psychological influence	<i>"If face to face, I would be very nervous, I will forget what I was planning to say..." (M9)</i>
The aim of school teaching is high score on examinations	<i>"...but there is one interesting phenomenon of school education, its goal is to gain higher scores, but you know, there is no direct relationship between the scores you get and your argumentation ability" (M8)</i>  <i>"...you know what school and the society is expected of us? High scores. We seldom have holiday because of this. We are expected to get higher scores by study, or more accurately, doing examination items repeatedly" (M7)</i>
Teacher's teaching schedule is tight	<i>"...I think it is helpful for improving my thinking ability, but in school learning, teachers taught more about the knowledge on the textbook because we have to take the examination, the teaching schedule is too tight to be completed, so SA seldom happens..." (M12)</i>  <i>"...we don't have much time to do this kind of activity, and school teachers and ourselves are busy at improving our scores" (M1)</i>
Fixed mindset	<i>"...well, I don't have any bad feelings (of taking the test), rather I think it's great to give us an opportunity to widen our horizon. Because normally, we do lots of examination</i>



	<p><i>items, and the way of our thinking has been a little fixed in a pattern. I feel good to go out of the zone of what we usually do” (F6)</i></p> <p><i>“...it is horrible that many students at our age have already had very strong mindsets...there are many representations of people’s mindset, like some of us are afraid of going out the comfort zone to be creative or to solve problems we never met before. Another example is that we started to do examination papers from an early age and we have done like hundreds and thousands of test papers, and all the questions have one answer, either right or wrong, when we met a question have a second answer, we are afraid to choose it even if we get it through reasoning...” (M8)</i></p>
<p>Ask for “how to get it” is painful than just memorize it</p>	<p><i>“...many schools they do not emphasis thinking, they care more on keep doing examination items and believe in short cuts. Just follow what teachers taught you, and to get scores by memorizing the formulas and conclusions that teachers have summarized for you. It is not difficult to get high score mostly by memory, teachers tell you the beginning and the ending of the story and leave apart what happened in the middle. In most cases, I think what happens in the middle is the argumentation process, I believe most students don’t know how to tell the whole story by adding the process. After all, it is much easier to memorize conclusion that to figure out how to get it.” (F13)</i></p> <p><i>“...it kept asking why, why, you know, I felt like breaking down to answer these ‘why’ questions. It is way more straight forward to throw several relevant formulas on the test paper...” (F5)</i></p>
<p>Care more about the final answer rather than the thinking process</p>	<p><i>“I don’t think I performed good since I didn’t even know the answer for several questions...” (M12)</i></p> <p><i>“...well, I chose the option of “none of the above” because he gives the wrong answer. I didn’t think too much on his argument after I found his answer is wrong, and I just choose “none of the above” (M8)</i></p>
<p>Applying tips to select multiple choice</p>	<p><i>“(laugh) actually, I feel that it always works</i></p>

items rather than think over it

*by choosing the positive answer if you feel what this person says is plausible, and you don't have to think over it very carefully" (M1)*

*"...like we usually do when taking examinations, if you find an option that is way too absolute, then this option must be wrong. It is always safe to choose relatively neutral one especially when you are not sure about the answer, that's what our teacher taught us" (M2)*

*"...yeah, I knew that. In fact, I thought this option might be right and I was hesitated whether to choose it. But I didn't dare to choose it for it is too absolute" (F13)*

## **Appendix 8 Participant information sheet for students (think aloud interview)**

### **Development and validation of assessment instrument for Chinese high school students' scientific argumentation competence in Physics.**

Dear Students,

I am Jinglu Zhang, a doctoral student from the School of Education at University of Bristol, UK. I am conducting a research study as part of the requirements of the degree of Doctor of Philosophy (PhD). I am designing a new kind of assessment, an assessment that measures students' ability to argue about science. The original idea of doing this study is that the new Physics curriculum puts new requirements on high school students' ability and our country are now trying to advance the new Gaokao reform that you are all familiar with. The results of this study are supposed to provide some advice on how to make Physics learning, teaching and assessment in line with the new curriculum requirements and the Gaokao reform. I am here to sincerely invite you to participate in this study. Your participation will help me to revise the test items and will finally contribute to the study. Your participation is appreciated. If you agree to participate in this study, please read the requirements of the research below.

- 1) This study will take you around 45 minutes to complete at school.
- 2) You will be invited to think aloud when you are doing part of a test (several test items) in the context of physics.
- 3) Before you starting to think aloud, the researcher will provide you some instructions on how to think aloud.
- 4) After you finished thinking aloud of the items, you will be invited for a follow-up interview, during which the researcher will ask you some questions related to your performance and the items, and you will be encouraged to talk about your feelings about doing the items.
- 5) All these stages will be audio recorded.
- 6) The interview will be transcribed for analysis, and the transcription will be shared with you for your approval of the data.

In addition, I would like to draw your attention to the following points:

- Participation in the study is voluntary

You do not have to be in this study if you do not want to. Your decision is fully respected. Your consent can be withdrawn within 10 days of the participation. No penalty is tied to withdrawal.

- Participation in the study is confidential

The data collected from you will be used only for research purpose. I will not record or use your name anywhere, so your identity will not be revealed.

- No risk is involved in your participation

There is no risk involved in your participation in this study. You are not being examined. Please do not worry if you cannot answer questions in the test since it has some difference with the test you usually take, your performance in this study is not related to your school assessment score. It is for research purpose only.

- The potential impact on your normal study schedule will be reduced to a minimum.

To minimize impact on your normal study, this study will be conducted when you are not taking any important exam or classes. Your teacher will select a time for the test that she/he believes to be most convenient.

If you are willing to participate in this study, you will receive a gift as reward, worth of around 1 British Pound (approximating 9 RMB).

If you have any questions or complaints regarding this study, please feel free to contact me at

jinglu.zhang@bristol.ac.uk, or my supervisor, Prof. William Browne at William.Browne@bristol.ac.uk. If you have any questions about your rights as a participant, please contact the University of Bristol ethics committee at gsoe-ethics@bristol.ac.uk.

Jinglu Zhang  
PhD Student  
School of Education  
Faculty of Social Science and Law  
University of Bristol

35 Berkeley Square,  
Bristol, BS8 1JA  
June 2020

## **Appendix 9 Participant information sheet for students (test and follow up interview)**

### **Development and validation of assessment instrument for Chinese high school students' scientific argumentation competence in Physics.**

Dear Students,

I am Jinglu Zhang, a doctoral student from the School of Education at University of Bristol, UK. I am conducting a research study as part of the requirements of the degree of Doctor of Philosophy (PhD). I am designing a new kind of assessment, an assessment that measures students' ability to argue about science. The original idea of doing this study is that the new Physics curriculum puts new requirements on high school students' ability and our country are now trying to advance the new Gaokao reform that you are all familiar with. The results of this study are supposed to provide some advice on how to make Physics learning, teaching and assessment in line with the new curriculum requirements and the Gaokao reform. You might be interested to take part in this new kind of Physics test, in which you are not being examined and judged about your achievement since the test cares more about your real thinking process and how you understand the questions. I am here to sincerely invite you to take part in the study and your participation that will contribute to the study is appreciated. If you agree to participate in this study, please read the requirements of the research below. This study will take you around 1-1.5 hours to complete at school. In more detail, you will be invited to do the following tasks:

- 1) Take a scientific argumentation (Physics) paper-pencil test lasting around 45-60 mins.
- 2) If you agree you might be interviewed based on your responses on the test to talk about your perspectives toward the test and your performance, the interview will last for about 30 mins and will be audio-recorded. There will be 1 out of 50 students be invited to the interview.
- 3) The interview will be transcribed for analysis, and the transcription will be shared with you for your approval of the data.
- 4) Your school test scores will be collected.

In addition, I would like to draw your attention to the following points:

- Participation in the study is voluntary

You do not have to be in this study if you do not want to. Your decision is fully respected. If you agree to take the test, you also have the right to choose whether you would like to participate in the interview or not. Your consent to the study can be withdrawn within 10 days of your participation. No penalty is tied to withdrawal.

- Participation in the study is confidential

The data collected from you will be used only for research purpose. I will not record or use your name anywhere, so your identity will not be revealed.

- No risk is involved in your participation

There is no risk involved in your participation in this study. You are not being examined. Please do not worry if you cannot answer questions in the test since it has some difference with the tests you usually take, your performance in this study is not related to your school assessment score. It is for research purpose only.

- The potential impact on your normal study schedule will be reduced to a minimum.

To minimise impact on your normal study, this study will be conducted when you are not taking any important exam or classes. Your teacher will select a time for the test that she/he believes to be most convenient.

If you are willing to participate in this study, you will receive a gift as reward, worth of around 20p (approximating 2 RMB).

If you have any questions or complaints regarding this study, please feel free to contact me at [jinglu.zhang@bristol.ac.uk](mailto:jinglu.zhang@bristol.ac.uk), or my supervisor, Prof. William Browne at

William.Browne@bristol.ac.uk. If you have any questions about your rights as a participant, please contact the University of Bristol ethics committee at gsoe-ethics@bristol.ac.uk.

Jinglu Zhang  
PhD Student  
School of Education  
Faculty of Social Science and Law  
University of Bristol

35 Berkeley Square,  
Bristol, BS8 1JA  
June 2020

## **Appendix 10 Participant information sheet for teachers**

### **Development and validation of assessment instrument for Chinese high school students' scientific argumentation competence in Physics.**

Dear Teachers,

I am Jinglu Zhang, a doctoral student from the School of Education at University of Bristol, UK. I am conducting a research study as part of the requirements of the degree of Doctor of Philosophy (PhD). This study aims to develop and validate an instrument to assess high school students' scientific argumentation competence. The original idea of doing this study is that the new Physics curriculum puts new requirements on high school students' ability and our country are now trying to advance the new Gaokao reform. Students will be invited to take a test in which they will be asked to explain and justify their claim when confronting Physics problems. Their performance on the test and in the interview will provide their attitude towards Physics learning and their strengths and weaknesses in arguing Physics problems. The right or wrong of their conclusion is not the focus of this study but their thinking process and their understanding toward Physics investigation and argumentation. I believe that after participating the study, some students if not all will gain better understanding of argumentation and the importance of evidence and justification in Physics. I am here to invite your students to participate in this study, and both you and your students' participation are appreciated.

If you agree that your students can be invited to participate in this study, please read the requirement of the research below. This study will take your students around 1-1.5 hours at the school. In more detail, some of your students will be invited to do a think aloud interview lasting about 45 mins (audio-recorded) and some students will be invited to do the following tasks:

- 1) Take a scientific argumentation (Physics) test lasting about 45-60 mins run as a paper-pencil test.
- 2) 1 out of 50 of them will be interviewed based on their responses on the test, the interview will last for about 30 mins and will be audio-recorded.
- 3) Your students' school test scores will be collected.

In addition, I would like to draw your attention to the following points:

- Participation in the study is voluntary

Your decision of approval of your students' participation is fully respected, and your consent of your students' participation can be withdrawn within 10 days of their participation. Your student does not have to be in this study if they do not want to as well. Your student's decision is fully respected. Your students can quit from the study within 10 days of their participation. No penalty is tied to withdrawal.

- Participation in the study is confidential

The data collected from your students will be used only for research purpose. The findings of the study may be published or presented at academic conferences, but yours and your students' identities and your school's identity will not be revealed.

- No risk is involved in your students' participation

There is no risk involved in your students' participation in this study. Your students might feel bad if they cannot answer the test questions, but they will be informed that the test has some difference with their school test, so they do not need to be frustrated. Your students' performance in this study is not related to their school assessment score. It is neither related to any evaluation of your teaching quality and your schools' management quality. It is for research purpose only.

- The potential impact on your normal study schedule will be reduced to the minimum. As the study will take place during school time, it may clash with some classes or activities

arranged by your school. To minimise such impact, this study will be conducted when students are not taking any important exam or classes. The specific time will be set based on agreement with you and your students.

If you allow your student to participate in this study, you will also receive a gift as reward, worth around 2 British pounds (approximating 19 RMB).

If you have any questions or complaints regarding this study, please feel free to contact me at [jinglu.zhang@bristol.ac.uk](mailto:jinglu.zhang@bristol.ac.uk), or my supervisor, Prof. William Browne at [William.Browne@bristol.ac.uk](mailto:William.Browne@bristol.ac.uk). If you have any questions about your rights as a participant, please contact the University of Bristol ethics committee at [gsoe-ethics@bristol.ac.uk](mailto:gsoe-ethics@bristol.ac.uk).

Jinglu Zhang  
PhD Student  
School of Education  
Faculty of Social Science and Law  
University of Bristol

35 Berkeley Square,  
Bristol, BS8 1JA  
June 2020



## Appendix 11 Students consent form for participation in research



School of Education  
35 Berkeley Square, Bristol  
BS8 1 JA  
[www.bristol.ac.uk](http://www.bristol.ac.uk)



China Scholarship Council, China

### **PART ONE: To be completed by the student researcher**

Name of student researcher: Jinglu Zhang

Contact: Phone number: +44 7529147728/+86 19953315216

Email: [jinglu.zhang@bristol.ac.uk](mailto:jinglu.zhang@bristol.ac.uk)

Name of supervisors: Professor William Browne, Dr Angeline Mbogo Barrett

Title of research project: Development and validation of an assessment instrument for Chinese high school students' scientific argumentation competence in Physics.

### **PART TWO: To be completed by participants**

I, \_\_\_\_\_ (*participant's name*), have been given and have read the Participant Information Form for Students describing the nature of the project being conducted by Jinglu Zhang for the research project entitled "Development and validation of an assessment instrument for Chinese high school students' scientific argumentation competence in Physics".

I understand the purpose and process of the research project and my involvement in it.

I also understand that

- I can withdraw my consent for my participation within 10 days of my participation without penalty, prejudice, negative consequences, repercussion, or disadvantage and demand that my personal data/information be permanently deleted from the researcher's records;
- The researcher will use the data and my personal information solely for this study;
- I will not be personally identified, and my personal data/information will remain confidential;
- The ethical aspects of the project have been approved by University of Bristol ethics committee.

If I have any questions about the research at any point in time, I will contact the researcher or the faculty supervisor.

**I hereby consent to my participation in the above research.**

Name of participant:

Signature:

Date:

## Appendix 12 Teachers consent form for participation in research



University of  
**BRISTOL**

School of Education

35 Berkeley Square, Bristol

China

Scholarship Council, China



BS8 1 JA

[www.bristol.ac.uk](http://www.bristol.ac.uk)

### **PART ONE: To be completed by the student researcher**

Name of student researcher: Jinglu Zhang

Contact: Phone number: +44 7529147728/+86 19953315216

Email: [jinglu.zhang@bristol.ac.uk](mailto:jinglu.zhang@bristol.ac.uk)

Name of supervisors: Professor William Browne, Dr Angeline Mbogo Barrett

Title of research project: Development and validation of an assessment instrument for Chinese high school students' scientific argumentation competence in Physics.

### **PART TWO: To be completed by participants' teacher**

I, \_\_\_\_\_ (*participant's teacher's name*), have been given and have read the Participant Information Form for Teachers describing the nature of the project being conducted by Jinglu Zhang for the research project entitled "Development and validation of an assessment instrument for Chinese high school students' scientific argumentation competence in Physics".

I understand the purpose and process of the research project and my students' involvement in it.

I also understand that

- I can withdraw my consent for my students' participation within 5 days of their participation without penalty, prejudice, negative consequences, repercussion, or disadvantage and demand that my students' personal data/information be permanently deleted from the researcher's records;
- The researcher will use the data and my students' personal information solely for this study;
- My students will not be personally identified, and my students' personal data/information will remain confidential;
- The ethical aspects of the project have been approved by University of Bristol ethics committee.

If I have any questions about the research at any point in time, I will contact the researcher or the faculty supervisor.

**I hereby consent to my participation in the above research.**

Name of participant:

Signature:

Date:

## Appendix 13 SoE research ethics form

It is important for members of the School of Education, as a community of researchers, to consider the ethical issues that arise, or may arise, in any research they propose to conduct. Increasingly, we are also accountable to external bodies to demonstrate that research proposals have had a degree of scrutiny. *This form must therefore be completed for each piece of research carried out by members of the School, both staff and students*

The SoE's process is designed to be supportive and educative. If you are preparing to submit a research proposal, you need to do the following:

**1. Complete the form on the back of this sheet**

A list of prompts for your discussion is given below. Not all these headings will be relevant for any particular proposal.

**2. Arrange a meeting with a fellow researcher**

The purpose of the meeting is to discuss ethical aspects of your proposed research, so you need to meet with someone with relevant research experience. Discussants are encouraged to take the role of critical friend and approach the research from the perspective of potential participants.

Track the changes in how your thinking has changed as a result of your decisions; this form is designed to act as a record of your discussion and any decisions you make.

**3. Upload a copy of this form and any other documents (e.g. information sheets, consent forms, materials) to the online ethics tool**

at: <https://dbms.ilrt.bris.ac.uk/red/ethics-online-tool/applications>.

**Please note: Following the upload you will need to answer ALL the questions on the ethics online survey and submit for approval by your supervisor (see the flowchart and user guides on the SoE Ethics Homepage).**

If you have any questions or queries, please contact the ethics co-ordinators at: [gsoe-ethics@bristol.ac.uk](mailto:gsoe-ethics@bristol.ac.uk)

**Please ensure that you allow time before any submission deadlines to complete this process.**

### Prompts for discussion

You are invited to consider the issues highlighted below and note any decisions made. You may wish to refer to relevant published ethical guidelines to prepare for your meeting. See <http://www.bris.ac.uk/education/research/networks/ethicscommittee/links/> for links to several such sets of guidelines.

1. Researcher access/exit
2. Power and participant relations
3. Information given to participants
4. Participant's right of withdrawal
5. Informed Consent
6. Complaints procedure
7. Safety and well-being of participants/researchers
8. Anonymity/confidentiality
9. Data collection
10. Data analysis
11. Data storage
12. Data protection (see: <http://www.bristol.ac.uk/secretary/data-protection/>)
13. Feedback
14. Responsibilities to colleagues/academic community
15. Reporting of research

Be aware that ethical responsibility continues throughout the research process. If further

issues arise as your research progresses, it may be appropriate to cycle again through the above process.

**Name(s):** Jinglu Zhang

**Proposed research project:** Development and validation of an assessment instrument for Chinese high school students' scientific argumentation competence in Physics.

**Proposed funder(s):** University of Bristol and China Scholarship Council

**Discussant for the ethics meeting:** Mr Dini Jiang

**Name of supervisor:** Professor William Browne, Dr Angeline Mbogo Barrett

**Has your supervisor seen this submitted draft of your ethics application?** Y

Please include an outline of the project or append a short (1 page) summary:

With the emphasis of science education shifting from the products and outcomes of learning to learning processes and practices, the significant role argumentation plays in science learning and research has been widely realized. Curriculums in China and internationally have been reformed to include scientific argumentation (SA) as an important component. However, assessments designed to measure students' ability to argue in a science context remain scarce especially in China. As a result, it is of great value to investigate valid ways to perform SA competence (SAC) assessment and draw a picture of Chinese students' current levels of SA ability. This study aims to develop and validate a SAC test in the context of Physics, during which process it will explore possible test influencing factors (content knowledge, item context and test scaffolds), as well as students' perspectives toward the test and their own performance to provide further implications for SAC assessment. In more detail, the assessment framework in this study, drawing on Toulmin (1958)'s argumentation pattern, Osborne's (2016) learning progression and Kuhn's (2013) idea of developing argument competence, constructs SAC from three dimensions (i.e. the identification, evaluation and production of argumentation).

This study uses a mixed-methods design with an iterative process of test development. Except for the preliminary pilot study, the second pilot test administration invites 2 Physics teachers to the item panel helping review the test items and scoring rubrics. Test items and scoring rubrics will be revised based on teachers' feedback, then 6 high school students aged around 17 will be invited for the think aloud interview and each of them will think aloud for around 45 mins. Test items and scoring rubrics will be revised again based on student's think aloud results. After that, 450 students from three high schools in one province in China will be invited to take the test and 8 of them for follow-up interview. Considering there are around 50 students in each class in China and each teacher teaches 2-3 classes, 1 or 2 physics teachers will be invited first to see if their students are willing to participate in the study. To complement the number of participants who are not willing to attend, other physics teachers will be contacted and invited. The test will last about 45-60 mins and the follow-up interview will be around 30 mins. The total data collection time for the second pilot will be around 6 weeks. The main study will invite 800 students from six high schools in two provinces in China to take the test with around 130 students be invited in each school. Later, 8 of them will be interviewed on their perspectives and experiences of taking the test. Their answers on the test will be scored, interviews will be audio recorded. The time costs for testing and interview will be around the same as in the second pilot and the total data collection time for the main study will be around 7 weeks. The two pilot studies aim to reveal possible problems with the test design to improve test quality. The results of the main study will indicate the quality of the test and students' current competence levels. Scores obtained from the main study will be analyzed to uncover the relationship between students' performance on the SAC test and their content knowledge, test context and test scaffolds. Interviews will be analyzed to obtain students' ideas and feelings toward taking the test.

Ethical issues discussed and decisions taken (see list of prompts overleaf):

### **1. Researcher access/exit**

The data collection will be conducted over two rounds. The first round will take place from June, 2020 to July, 2020, and the second round will start from October, 2020 and end in December, 2020. Considering the scale of the research, this study will take place at several high schools in two provinces in China. The schools will be contacted via my friends or colleagues who are physics teachers. Due to my previous study experience, most of my friends and colleagues are physics teachers whose workplace located in different cities in China. With their support, I will have access to the students. Among the two provinces, one of which is my hometown, and another is also not far away from my home. Many of my colleagues and classmates are teaching physics in the two provinces, which will secure my access to these schools.

The planned data collection date above has taken the situation of coronavirus into consideration (Relevant experts predict that the coronavirus in China is supposed to be totally controlled before the end of April, so June seems like an appropriate time to collect data). I also confirmed with high school teachers and they said it sounds reasonable to collect data in June. However, if the situation in China has not gotten better at that time, the two rounds of data collection will be delayed and rescheduled to avoid the possible safety risk.

Considering this study is in line with the new physics curriculum document in China, teachers and students might be quite interested in the study. I have contacted some of the teachers and they are glad to participate in the study. In addition, the findings and implications for physics learning, teaching and assessment will be shared with these teachers and students after finishing the research.

### **2. Power and participant relations**

Despite the volunteer nature of the study and the supposed equal relationship between participants and researcher, students might still take the researcher (me) as the one that has power since I am their teacher's friend. So, teachers will be fully informed of the importance of respecting students' decision and not forcing them to participate. Before each data collection stage, I will double check with students about the volunteer nature of the study and that their decision will be fully respected.

### **3. Information given to participants**

Since the age of participants in this study is around 17, their parents need not be required to give consent on behalf of their children. Consent will be obtained from teachers and students, and they will be informed of all possible issues that might influence their decision of participation including the aim, methods and potential consequences of the study. Details can be seen in the information sheets and consent forms. Specifically, students who take part in the think aloud protocols and those who take the test will be given a consent form and information sheet. Teachers will be informed of the specific design and content of the test to check whether it might influence their normal teaching. Except for the consent form and information sheet, students will be informed orally of the research aims, the importance of their participation, their right to withdraw and the confidentiality before each data collection stage.

### **4. Participant's right of withdrawal**

The voluntary nature and the importance of participants' willingness of participation will be sufficiently informed to students and teachers. Considering China's learning and teaching situation, the test will take place in a physics course or self-study course, students will be fully informed that if they are not willing to participate in the study they can do their own normal study as usual. Besides, students will be informed of their right on choosing whether to participate in this study and given the ability to withdraw from the study within 10 days of their participation. Teachers will also be given full command on their decision of whether they would like their students to take part in the study or not. Teachers' consent of their

students' participation of the study can be withdrawn within 5 days of students' participation.

#### **5. Informed Consent**

Teachers and students who take part in the think aloud interview and who participate in the test and follow-up interview will respectively receive a written consent form about participation, which includes the research aims, methods, the potential consequences, the right of withdrawal, the confidentiality and the reward for participation. All the participants will be informed of the above information orally before each data collection stage, and their oral agreement will be obtained as well to confirm again that they are willing to participate in the study and they have understand all the requirements of participating in the study.

#### **6. Complaints procedure**

They will be informed of the right to complain and the complaint procedures in the consent form. The contact information of my supervisor(s) will be provided in the consent form.

#### **7. Safety and well-being of participants/researchers**

All the stages of data collection will take place in the school of the students which is a familiar setting. The think aloud interview, and follow-up interview will take place in a room at school and the test will take place in the participants' classroom. Besides, all the data collection procedures will be conducted during times when the school is open. Teachers will be informed of the exact time and place of the interview, and the time for test will be agreed by teachers based on their normal schedule. So, the safety of students will be fully guaranteed. As for participants' well-being, students and teachers will be informed that their performance will not be related to any form of school assessment and will not be used as any indicator to evaluate the quality of teaching or the school. The data is used for research purpose only. Taken together, taking part in this research should not impact on participants' welfare but if they are affected by taking part, they can seek support at the welfare office in their school.

In order to minimize the influence on the normal schedule of the school, data collection will be conducted when there is no important activities e.g. major exams and the procedure of data collection will be designed to save participants' time. For each data collection stage, the time and place will be discussed with teachers and students to minimize influences on them. As for the researcher's (my) safety and wellbeing, considering I will collect data in different schools, I will contact teachers beforehand to arrange the data collection schedule to save my own time and cost. Since the choice of cities also considered my familiarity with them and whether I have family member or friends in that city, my safety and wellbeing risks will be minimized.

#### **8. Anonymity/confidentiality**

In the think aloud interview, students' name will not be collected. In the test, students' name will be collected, while students have the right to not provide their name. The reason for collecting their name is that in later data analysis, the relationship of their performance on the test and on a previous school test (collected directly from teachers) will be analysed, and the follow-up interview will be conducted based on their test performance. So, in the first round of data collection that does not need to collect students' school test scores, their names will be removed from the data after the follow-up interview. In the main study, their names will be removed after matching their school test with the test in this study. After removing their names, all the data analysis and reports in later procedures will be anonymised. No participants' personal information will be revealed in any form of report of this study. Students will be fully informed of the process above and will choose whether they provide their name or not.

#### **9. Data collection**

The data to be collected includes:

- a) Teachers' panel review of the test (remotely or face to face; audio-recorded and noted

- during the panel);
- b) Students' think aloud interview of items (face to face; audio-recorded; within 40 mins);
- c) Students' responses on the test paper (paper-pencil test; within 1hr);
- d) Students' follow-up interview based on their performance on the test (face to face; audio-recorded; within 40 mins);
- e) Students' previous school test that can best represent their content knowledge proficiency.

Before each data collection stage, participants' willingness to participate and their awareness of the research will be confirmed again. Considering students might not try their best to do the test since it is not related to any kind of school assessment, I will try to convey the information to them that their serious participation is very important and meaningful. Both my mobile phone and a voice recorder will be used to record interviews to ensure the success of data collection and storage.

### **10. Data analysis**

All the data will be analyzed in a fair, lawful and transparent way:

- a) Students' think aloud interview of items will be transcribed and checked by another colleague, and the transcription will be shared with participants for agreement of accuracy. Then the transcription will be coded by the researcher and analysed qualitatively using NVivo to help revise the items.
- b) Students' responses on the test will be scored based on the rubrics by two raters, one physics teacher who is familiar with the study and the researcher. Scores will be analysed quantitatively using the R package.
- c) The follow-up interview will be transcribed, and the accuracy of transcription will be checked by another researcher. Then the transcription will be shared with the participants for agreement of accuracy. The transcript will be analysed with thematic analysis on the NVivo platform.
- d) The relationship between students' school test and SAC test will be analysed via SPSS.

### **11. Data storage**

All the data will be stored under safe conditions and regularly backed up on the university platform and my own laptop, both of which are password protected. Before students' names are removed from the data, all the data will be encrypted. No other people could get access to the data except for the researcher and supervisors. All the conducts in data storage and data processing will abide by the UK Data Protection Act and the General Data Protection Regulation. Besides, all the data collected in this study will only be stored for the duration of the research.

### **12. Data protection**

All the conducts in data storage and data processing will abide by the UK Data Protection Act and the General Data Protection Regulation. None of the data will be shared with any third party.

### **13. Feedback**

During the final phase of my PhD study, the research findings and implications for SAC learning, teaching and assessment in Physics will be shared with students and teachers who participated in the study upon request.

### **14. Responsibilities to colleagues/academic community**

I will share my research design, findings and implications, with integrity, responsibility and high professional standard, with my colleagues, researchers, SAC assessment developers, my scholarship sponsors (CSC) and any relevant or interested stakeholders. The goal is to connect academia with the public, and to transform the research into practice that will bring

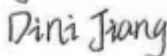
changes to science education.

### **15. Reporting of research**

The research will be firstly reported in the form of my doctoral dissertation. If possible, it will also be presented at academic conferences and seminars, journal articles or book chapters. All the results from the study will be reported honestly that conforms to the academic integrity.

If you feel you need to discuss any issue further, or to highlight difficulties, please contact the GSoE's ethics co-ordinators who will suggest possible ways forward.

Signed:  (Researcher) Jinglu Zhang

Signed:  (Discussant) Dini Jiang

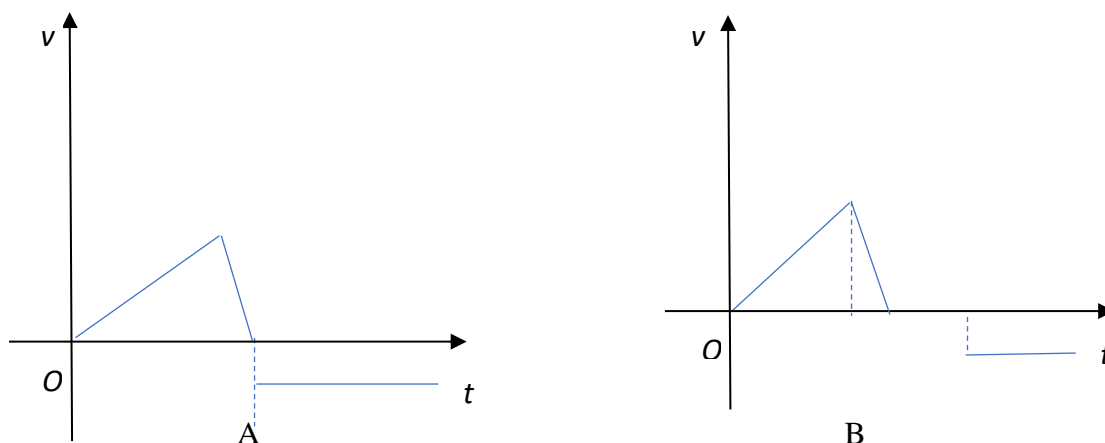
Date: 26<sup>th</sup> March, 2020



## Appendix 14 Test version I-teacher

### Task 1. Identification

After receiving a delivery phone call, Xiao Li rushed to the pickup point and then walked home after picking up a package. There is a straight road between the delivery point and home. A and B are possible  $v$ - $t$  images of Xiao Li's movement from going to pick up the package to returning home.



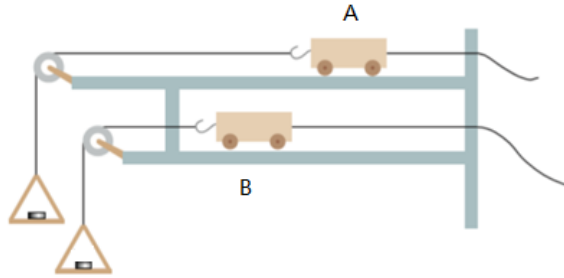
Xiao Li agrees with the process drawn by Figure B:

- (1) The first picture does not indicate the movement status of waiting for pickup.
- (2) The beginning of the second graph is a straight line with a positive slope, and then a straight line with a negative slope. Both are on the positive semi-axis of the  $y$ -axis. The next section coincides with the horizontal axis, and the last horizontal straight line is on the negative semi-axis of the  $y$ -axis.
- (3) Xiao Li runs for the delivery, so speed up the movement, and decelerate before reaching the pickup point. Walk home, so the speed will be smaller, and the direction of movement is the opposite. In the  $v$ - $t$  image, a straight line with a positive slope represents uniform acceleration, a negative slope means uniform deceleration, a horizontal straight line means uniform speed, and coincidence with the  $X$  axis means that the speed is 0.
- (4) The second picture is the most reasonable.

In the above description, Xiao Li's claim is       (4)      . The evidence Xiao Li used to support his claim is       (2)      . Xiao Li's reason for using the evidence is       (3)      . Xiao Li's rebuttal towards the first picture is       (1)      .

### Task 2. Evaluation-use of evidence

As shown in the figure, the two trolleys A and B start to move from standstill under the pulling force of the groove plate, and the two cars move for the same time (smooth track, smooth pulley). Three students a, b and c each give evidence to prove the conclusion that "the mass of the slot code in slot B is greater than the slot code in slot A".



A: Car A uses more environmentally friendly materials than Car B. Car B uses thicker ropes than Car A.

B: Both cars have the same mass, and the displacement of A is greater than B.

C: Both cars passed the same displacement.

Student d thinks that the evidence given by the 3 other students cannot reach a conclusion.

Please enter the reason why evidence given by abc cannot support the conclusion.

A ----- reason: ( 2 ); B ----- reason: ( 1 ); C ----- reason: ( 3 )

- (1) The evidence contradicts the conclusions.
- (2) The evidence is not relevant to the conclusions.
- (3) The evidence is insufficient.

### Task 3. Production-explanation

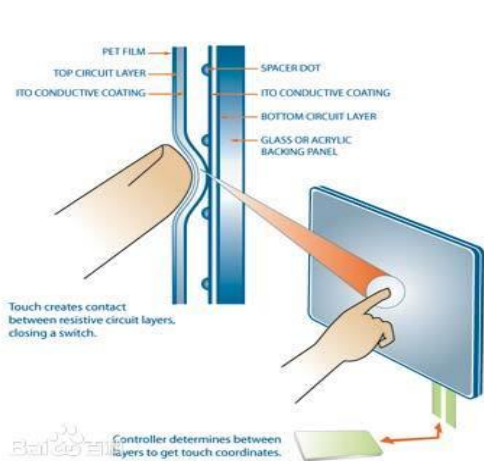
The table below recorded the weight of a student (weighing 55kg on the ground) at four moments when taking an elevator.

t1	t2	t3	t4
55kg	50kg	55kg	60kg

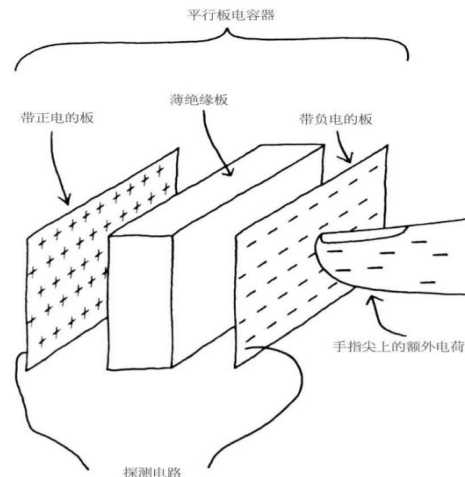
What kind of movement might the elevator undergo? Why? (Please describe the possible movement of the elevator at each time point and time period and show evidence and reasons to fully support your description).

### Task 4. Production-use of evidence; explanation; rebuttal

There are usually two types of touch screens, one is the resistive touch screen (A), and the other is capacitive touch screen (B). The resistive touch screen relies on the pressure generated when the screen is clicked, so that the conductive layers under the screen contact each other to form a closed circuit, and the internal chip records the position where the pressure is generated to determine the clicked position. Capacitive touch screens rely on the conductivity of the human body. When touching, the human body and the electrode plate under the touch screen form a capacitance. By recording the position where the current changes, the position of the touch point can be determined.



A



B

- Which screen (s) responds when tapping on the screen while wearing dry cotton gloves? Which one (s) of the following statements support your conclusion? (bcd)
  - Cotton gloves can conduct electricity
  - The human body and the plates under the capacitive screen form a capacitor
  - The conductive layers under the resistive touch screen contact each other to form a closed circuit
  - Cotton gloves are not conductive
- Why did your chosen narrative support your conclusion? (Please describe causality accurately and comprehensively.)
- Xiaohong and Xiaolan found that when there was water on the capacitive screen, the reaction would fail. The two started a discussion.



Water prevents fingers from directly touching the screen, capacitors cannot be formed where water is present, and current does not change. As a result, the response failed.

Xiaohong

I do not think so.



Xiaolan

How do you think Xiaolan will respond to Xiaohong? (Please clearly point out Xiaohong's mistakes, explain why they are wrong, then give your own opinion and explain why your opinion is correct.)

### Task 5. Production-rebuttal

Hulk and Superman each take a weight scale against each other, neither of them backed up, so whose weight scale has larger indication?



It must be the number in Superman's hand is large, because Hulk is strong, and the thrust on Superman's weight scale is larger.

A



Hulk's indication is larger, because Hulk is strong and pushes his weight scale in more strength.

B

Who do you agree with? why? (Indicate the points you agree or disagree with and explain why. Then give your own points and explanation.)

### Task 6 Evaluation-Explanation

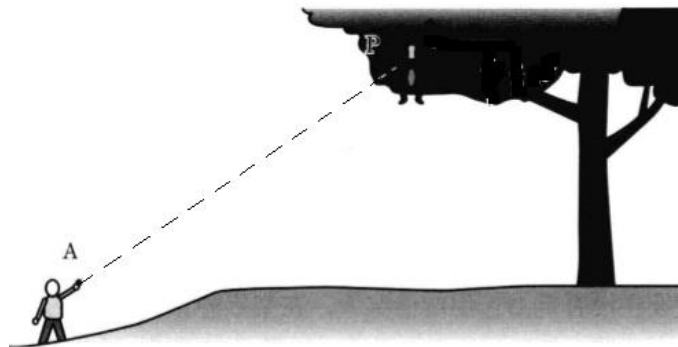
Lee stands at point A and aims at the monkey (at point P) in the tree with his marbles. When the monkey saw the marble firing, it released its hands that hold the branch, but it is still hit by the marble. After discussing what happened, Lee's friends gave the following explanation.

- a. The monkey is hit because the marble fired fast.
- b. The monkey is hit because both marble and the monkey are under gravity.
- c. The marble hit the monkey because they were under gravity and the marble did not move at a uniform linear speed.

Lee does not support these three explanations. What do you think is the reason?

a --- reason ( 1 )      b --- reason ( 2 )      c --- reason ( 2 )

(1) Causality is not established. (2) The explanation of causality is incoherent and inadequate.



## Appendix 15 Scoring rubrics I

<i>Task 3: Production</i> <i>Item 3-P-EX: Generating explanation</i>	
<i>Score</i>	<i>Description</i>
5	Student uses coherent articulation to explain the process with correct causal relationship.
4	Student provides correct causal relationships and understand why this phenomenon happens without explaining the process or explains it not fully correct.
3	Student tries to provide explanation, but conclusions or causal relationships are incorrect.
2	Student does not give explanation but provides correct conclusion to the process of movement.
1	Student does not give explanation but provides partly correct conclusion to the process of movement.
0	Student does not give explanation and does not provide correct conclusion to any points or stage of movement.

<i>Task 4: Production</i> <i>Item 4-1-P-UE: use of evidence.</i>	
<i>Score</i>	<i>Description</i>
2	Student selects A, and select b,c,d.
1	Student selects A but does not include all relevant evidence.
0	Student Selects B or student selects A but include incorrect evidence a.

<i>Task 4: Production</i> <i>Item 4-2-P-EX: Generating explanation</i>	
<i>Score</i>	<i>Description</i>
3	Student uses coherent articulation to explain the process with correct causal relationship.
2	Student provides correct causal relationships without explain the causality or explain it not fully correct.
1	Student tries to provide explanation, but conclusions or causal relationships are incorrect.
0	Student does not give explanation or repeats statements in item 4-1

<i>Task 4: Production</i> <i>Item 4-3-P-R: Rebuttal</i>	
<i>Score</i>	<i>Description</i>
5	Student points out the mistake and explain why it is wrong and provides his/her own claim and justifies it correctly.
4	Student points out the mistake and explain why it is wrong and provide his/her own claim but justifies it incorrectly.
3	Student points out the mistake and explain why it is wrong and provide his/her

	own claim without justification.
2	Student points out the mistake correctly and explain why it is wrong without providing his/her own claim or providing wrong claim.
1	Student points out the mistake correctly without explaining why.
0	Student does not point out where the mistake is or does not recognize the mistake correctly or provides irrelevant statement.

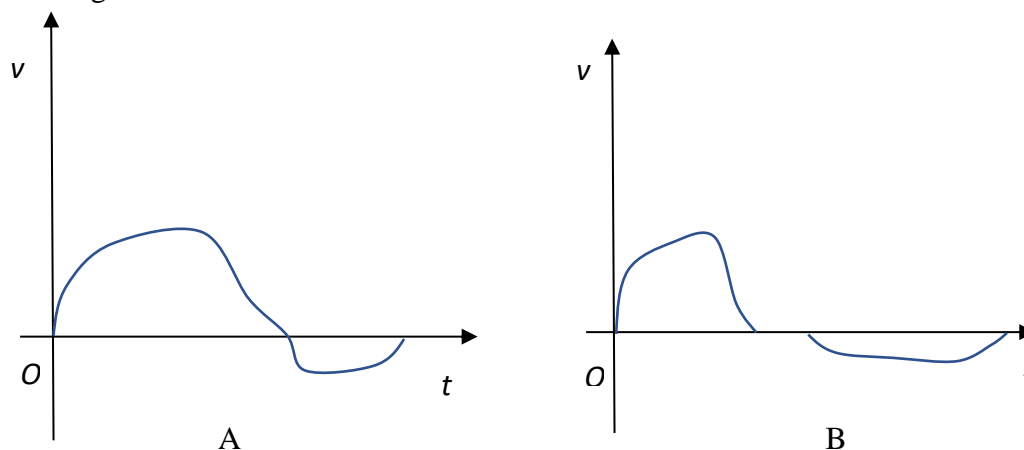
*Task 5: Production*  
*Item 5-P-R: Rebuttal*

<i>Score</i>	<i>Description</i>
5	Student points out the mistake and explains why it is wrong and provides his/her own claim and justifies correctly.
4	Student points out the mistake and explains why it is wrong and provides his/her own claim but justifies incorrectly.
3	Student points out the mistake and explains why it is wrong and provides his/her own claim without justification.
2	Student points out the mistake correctly and explain why it is wrong without providing his/her own claim or provides wrong claim.
1	Student points out the mistake correctly without explaining why.
0	Student does not point out where the mistake is or does not recognize the mistake correctly or provides irrelevant statement.

## Appendix 16 Test version I-students

### Task 1.

After receiving a delivery phone call, Xiao Li rushed to the pickup point and then walked home after picking up a package. There is a straight road between the delivery point and home. A and B are possible  $v$ - $t$  images of Xiao Li's movement from going to pick up the package to returning home.



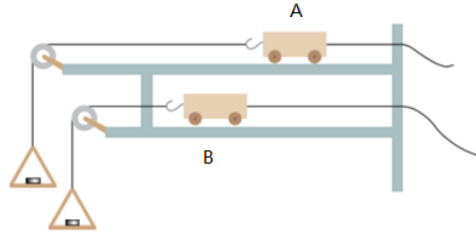
Xiao Li agrees with the process drawn by Figure B:

- (1) The average value of  $v$  is greater when  $v$  is positive than when  $v$  is negative, so it should not take longer when  $v$  is positive than when  $v$  is negative. The  $v$  in graph A changes from a positive value to a negative value directly, there is no state indicating waiting for pickup.
- (2) The average value of  $v$  on the positive half-axis of the y-axis in Figure B is larger, the next section coincides with the horizontal axis, and the last section is on the negative half-axis of the y-axis, and the  $v$  value is small. Figure B shows a shorter  $t$  when  $v$  is positive and a longer  $t$  when  $v$  is negative.
- (3) Xiao Li runs for the delivery, so speed up the movement, and decelerate before reaching the pickup point. Walking home, so the speed will be smaller, and the direction of movement is the opposite. Therefore, it takes less time to go to the delivery point than to go home. The image coincident with the X-axis means  $v$  is 0, representing the time to wait for pickup.
- (4) The second picture is more reasonable.

In the above description, Xiao Li's claim is 4. The evidence Xiao Li used to support his claim is 2. Xiao Li's reason for using the evidence is 1. Xiao Li's rebuttal towards the first picture is 3. (Please fill in the corresponding serial number in the horizontal line)

### Task 2.

As shown in the figure, the two trolleys A and B start to move from standstill under the pulling force of the groove plate, and both track and pulley are smooth. Three students a, b and c each give evidence to prove the conclusion that "the mass of the slot code in slot B is greater than the slot code in slot A".



- a: Car A uses more environmentally friendly materials than Car B. Car B uses thicker ropes than Car A.
- b: Both cars have the same mass, and the displacement of B is greater than A.
- c: The two cars have the same mass, and the acceleration of A is greater than that of B.

Student d thinks that the evidence given by the 3 other students cannot reach the conclusion. Please enter the reason why evidence given by abc cannot support the conclusion.

A ----- reason: ( 2 ); B ----- reason: ( 1 ); C ----- reason: ( 3 )

- (1) The evidence contradicts the conclusions.
- (2) The evidence is not relevant to the conclusions.
- (3) The evidence is insufficient.

**Task 3.**

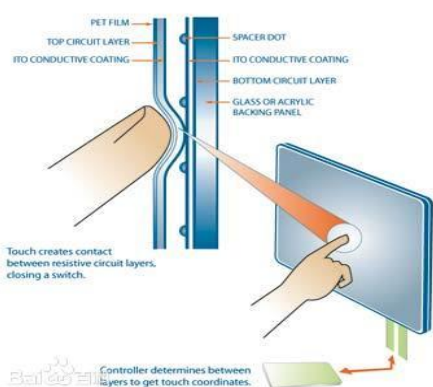
The table below recorded the weight of a student (weighing 55kg on the ground) at four moments when taking an elevator.

t1	t2	t3	t4
55kg	50kg	55kg	60kg

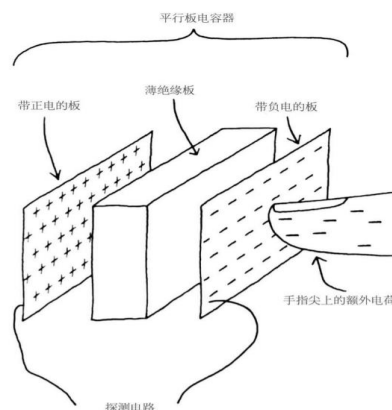
What kind of movement might the elevator undergo? Why?

**Task 4.**

There are usually two types of touch screens, one is the resistive touch screen (A), and the other is capacitive touch screen (B). The resistive touch screen relies on the pressure generated when the screen is clicked, so that the conductive layers under the screen contact each other to form a closed circuit, and the internal chip records the position where the pressure is generated to determine the clicked position. Capacitive touch screens rely on the conductivity of the human body. When touching, the human body and the electrode plate under the touch screen form a capacitance. By recording the position where the current changes, the position of the touch point can be determined.



A



B

1. Which screen(s) responds when tapping on the screen while wearing dry cotton gloves?



Which one (s) of the following statements support your conclusion? ( bcd)

- a. Cotton gloves can conduct electricity
- b. The human body can conduct electricity
- c. Capacitors have two plates
- d. Cotton gloves are not conductive
- e. Current is generated when the amount of charge on the capacitor plate changes

2. Why did your chosen information support your conclusion?

3. Xiaohong and Xiaolan found that when there was water on the capacitive screen, the reaction would fail. The two started a discussion.

Xiaohong: Water prevents fingers from directly touching the screen, capacitors cannot be formed where water is present, and current does not change. As a result, the response failed.

Xiaolan: I do not think so.

How do you think Xiaolan will respond to Xiaohong?

### Task 5.

Hulk and Superman each takes a weight scale against each other, neither of them backed up, so whose weight scale has larger indication?

A: It must be the number in Superman's hand is large, because Hulk is strong, and the thrust on Superman's weight scale is larger.

B: Hulk's indication is larger, because Hulk is strong and pushes his weight scale in more strength.

Who do you agree with? why?

### Task 6

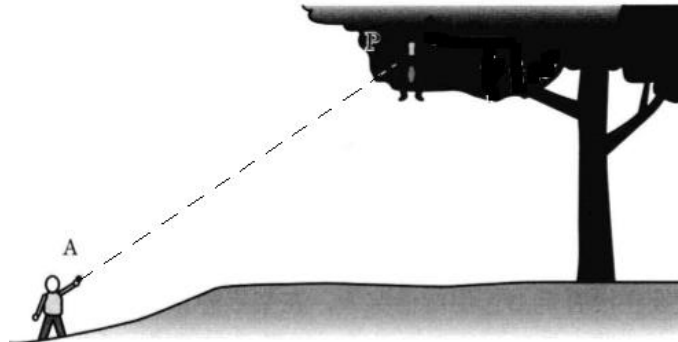
Lee stands at point A and aims at the monkey (at point P) in the tree with his marbles. When the monkey saw the marble firing, it released its hands that hold the branch, but it is still hit by the marble. After discussing what happened, Lee's friends gave the following explanation.

- a. The monkey is hit because the marble fired fast.
- b. The monkey is hit because both marble and the monkey are under gravity.
- c. The marble hit the monkey because they were under gravity and the marble did not move at a uniform linear speed.

Lee does not support these three explanations. What do you think is the reason?

a --- reason ( 1 )      b --- reason ( 2 )      c --- reason ( 2 )

(1) Causality is not established. (2) The explanation of causality is incoherent and inadequate.



## Appendix 17 Test version II-teachers

### First draft of SAC test (with scaffold)

Scaffold: The scaffold is helping you get a sense of what is claim, evidence, reason and rebuttal, instead of showing what is high-quality argument.

*Example:* Ball A has bigger inertia. Ball A has a mass of 10kg, and ball B has a mass of 5kg. The larger the mass, the greater the inertia. You said that the inertia of ball B is bigger because its high velocity, which is not right. Inertia has nothing to do with velocity.

**Claim:** Ball A has bigger inertia

**Evidence** (which can be used to support the claim): Ball A has a mass of 10kg, and ball B has a mass of 5kg

**Reason** (which explains why certain evidence can support the claim): The larger the mass, the greater the inertia

**Rebuttal** (which point out whether an argument is good or not and why): You said that the inertia of ball B is bigger because its high velocity is not right, inertia has nothing to do with velocity.

**The teacher led Bob and Jane to Y university for a summer camp. What they saw and heard on the way could cause so much thinking and discussion about physics! Let's go and have a look!**

**Scene 1: It's hard to take a taxi on a rainy day. The taxi driver is driving slower although they are about to miss the train.**



*Bob:* I think water reduces friction between objects. On a rainy road, the speed decreases. Since the water between the tires and the road reduces friction and makes the braking distance longer.

*Jane:* I think water adds friction, like when you're counting money, it's easier to separate the money by dipping it in your finger. In rainy days, the vehicle becomes slow is not because of the smaller friction, but because the sight is not clear, affecting the driver to evaluate the road condition.

1. What is Bob's claim? *water reduces friction*  
A. There is no claim. B His claim is \_\_\_\_\_
2. What is Bob's evidence? *On a rainy road, the speed decreases*  
A. There is no evidence. B. His evidence is (mark off use "\_\_\_\_\_")
3. What is Bob's reason? *road reduces friction and makes the braking distance longer*  
A. There is no reason. B. His reason is (mark off use "-----")
4. Which sentence is rebuttal? *I think water adds friction..... affecting the driver to*

evaluate the road condition.

A. There is no rebuttal. B. The rebuttal is (mark off use “  ”)

Later, they got the following experimental data from internet.

Table 1 maximum static friction between shoes and the ground

	Shoes	First (N)	Second (N)	Third (N)	Average (N)
Asphalt surface	Dry	2.84	2.84	2.84	2.84
	Wet	2.35	2.25	2.45	2.35
Cement floor	Dry	2.84	2.74	2.79	2.79
	Wet	2.06	1.86	1.86	1.93
Terrazzo floor	Dry	1.72	1.76	1.76	1.75
	Wet	1.86	1.91	1.91	1.90

Table 2 maximum static friction between leather and paper currency

	First (N)	Second (N)	Third (N)	Average (N)
Dry	0.93	0.93	0.93	0.93
One water spray	1.37	1.42	1.27	1.35
Two water spray	1.47	1.37	1.47	1.44
Three water spray	1.37	1.27	1.32	1.32

Table 3. Maximum static friction between sandpaper

	Water volume	Average value (N)	Sandpaper with waterproof spray	Water volume	Average value(N)
	Ordinary sandpaper	Dry		1.1	Dry
One spray		1.25	One spray	1.25	
Two sprays		1	Two sprays	1.16	
Three sprays		0.98	Three sprays	1.04	

Bob: Water sometimes increases friction and sometimes decreases it because of the different materials on the contact surface. The friction between soles and asphalt and cement floor decreases after water is added, while between soles and terrazzo floor increases. This difference follows difference on surface material.

Jane: I disagree. The experimental data showed that the friction between the leather and bill increased when a small amount of water was added, and then decreased. So, the influence of water on friction is not related to the material but to the amount of water.

- What is Bob’s claim? *The effect of water on friction is related to the material*  
A. There is no claim. B His claim is \_\_\_\_\_
- What is Bob’s evidence? *The friction of soles and asphalt, cement floor decreases after water is added, with soles and terrazzo floor increases*  
A. There is no evidence. B. His evidence is (mark off use “ \_\_\_\_\_ ”)
- If so, which of the following do you think is true of his evidence? (can choose more than one option) **A**  
A. The evidence and claim are relevant.

- B. The evidence is sufficient.  
C. None of above
8. What is Bob's reason? *This difference follows difference on surface material*  
A. There is no reason. B. His reason is (mark off use "-----")
9. If so, which of the following do you think is true of the reason? (can choose more than one) *A B*  
A. Reasons are related to evidence and claim  
B. Reasons are reasonable  
C. Reasons are comprehensive  
D. None of above
10. Which sentence is rebuttal? *I disagree...the amount of water*  
A. There is no rebuttal. B. The rebuttal is (mark off use "~~~~~")
11. If so, which of the following do you think is true of the rebuttal? (can choose more than one) *A B*  
A. Accurately pointed out the other's mistake  
B. Rebuttal based on appropriate evidence  
C. Rebuttal is reasonable  
D. None of above
12. Who do you agree with more?  
A. Bob B. Jane C. My own opinion\_\_\_\_\_
13. What evidence makes you agree with Bob/Jane/Yourself?
14. How does the evidence support Bob/Jane/yourself?
15. Why do you think Jane/Bob/both is wrong?

**Scene 2: They finally catch the train before it leaves. There were several children in the carriage, and one of them flew his toy helicopter.**



*Bob: It's dangerous to play a helicopter in the carriage.*

*Jane: Why?*

*Bob: The helicopter will not continue to hover steadily after driving. It will bump into people or the door of the carriage.*

*Jane: No, neither we nor our bags slid back after driving. The movement of stuff in the carriage and the carriage are the same.*

1. What is Jane's claim? *It is not dangerous to play a helicopter on the train*  
A. There is no claim. B Her claim is\_\_\_\_\_
2. What is Jane's evidence? *Neither we nor our bags slid back after driving*  
A. There is no evidence. B. Her evidence is (mark off use "\_\_\_\_\_")
3. If so, which of the following do you think is true of her evidence? (can choose more than one) *A*  
A. The evidence and claim are relevant.

- B. The evidence is sufficient.  
C. None of above
4. What is Jane's reason? *The movement of stuff in the carriage and the carriage are the same*  
A. There is no reason. B. Her reason is (mark off use "-----")
5. If so, which of the following do you think is true of Jane's reason? (can choose more than one) *A*  
A. Reasons are related to evidence and claim  
B. Reasons are reasonable  
C. Reasons are comprehensive  
D. None of above
6. Which sentence is rebuttal? *No.... the same*  
A. There is no rebuttal. B. The rebuttal is (mark off use "~~~~~")
7. If so, which of the following do you think is true of the rebuttal? (can choose more than one) *D*  
A. Accurately pointed out the other's mistake  
B. Rebuttal based on appropriate evidence  
C. Rebuttal is reasonable  
D. None of above
8. Who do you agree with more?  
A. Bob B. Jane C. My own opinion \_\_\_\_\_
9. Which of the following facts/theories support Bob/Jane/yourself? *A, C*  
A. Bodies that are not subjected to external forces have a tendency to remain in motion  
B. When a train is traveling in a straight line at a constant speed, the object in the carriage is traveling at the same speed as the train  
C. The train accelerates when it starts  
D. People lean back when the train is speeding up  
E. when the train accelerates, the front carriage drives the rear carriage
10. How does the evidence support Bob/Jane/yourself?
11. Why do you think Jane/Bob/both is wrong?

**Scene 3: The train passed by an amusement park, and they saw many people riding the roller coaster.**

*Bob: On the top side, the speed of the coaster should be no less than  $\sqrt{gR}$ ,  $R$  is the radius of the circular track.*

*Jane: This is not accurate.*

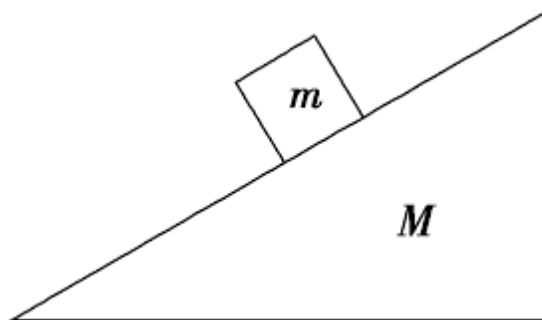
1. Which of the following statements could support Bob's claim? *A*  
A. At the top, the gravity and the orbital's supporting force go in the same direction  
B. When the roller coaster has the smallest speed, only gravity provides centripetal force  
C. The circular orbit has a large radius  
D. None of the above
2. How does the statement support(s) Bob?
3. Do you believe in what Jane said?  
A. yes B. no
4. Why do you believe/not believe in Jane?

**Scene 4: They finally arrived at University Y and went to the physics laboratory on the first day.**

They were excited to see all the experimental equipment. The first idea was to measure the kinetic friction factor ( $\mu$ ) they had just learned. Bob found the following equipment: A. Angle adjustable bevel B. Wooden block C. Balance D. Protractor

*Bob: The balance can be used to obtain the block's mass  $m$ , and the protractor can get bevel angle  $\theta$ . Adjusting the bevel until the block is static on it, then  $F_N = mg \cos \theta$ ,  $F_f = mg \sin \theta$ , according to  $F_f = \mu F_N$ , we can get  $\mu$ .*

*Jane: Your method is not right. Wood plank and spring dynamometer are needed.*



1. Which of the following is true about Bob's evidence? (can choose more than one) **A**
  - A. The evidence and claim are relevant.
  - B. The evidence is sufficient.
  - C. None of above
2. Which of the following is true about Bob's reason? (can choose more than one) **A**
  - A. Reasons are related to evidence and claim
  - B. Reasons are reasonable
  - C. Reasons are comprehensive
  - D. None of above
3. What is Jane's claim? *Bob's method is wrong*
  - A. There is no claim. B Her claim is \_\_\_\_\_
4. What is Jane's evidence? **A**
  - A. There is no evidence. B. Her evidence is (mark off use "\_\_\_\_\_")
5. What is Jane's reason? **A**
  - A. There is no reason. B. Her reason is (mark off use "-----")
6. Which sentence is rebuttal? *Your method is not right. Wood plank and spring dynamometer are needed*
  - A. There is no rebuttal. B. The rebuttal is (mark off use "\_\_\_\_\_")
7. If so, which of the following do you think is true of the rebuttal? (can choose more than one) **C**
  - A. Accurately pointed out the other's mistake
  - B. Rebuttal based on appropriate evidence
  - C. Rebuttal is reasonable
  - D. None of above

**Scene 5: Later, they measured the mass of the car on an existing device in the laboratory and recorded the following information. Track A and B are the same.**

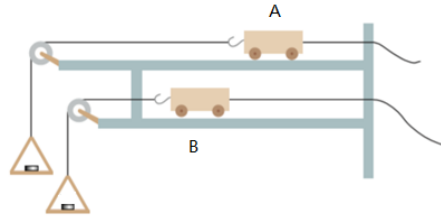


Table 3

Track	Initial speed $V_0$ (m/s)	Displacement $x$ (m)	Diameter of rope (m)	Mass of weight (g)	Time of motion (s)	Car material
A	0	0.2	0.04	20	4	Non-Eco-friendly
B	0	0.15	0.02	30	2	Eco-friendly

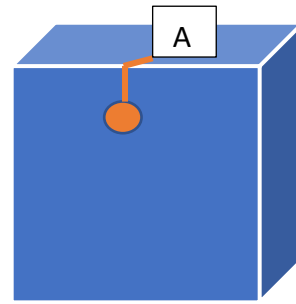
Bob: The mass of car A is larger. The material of car A is not eco-friendly, and eco-friendly materials are usually lighter.

- What is Bob's claim? *The mass of car A is larger.*  
A. There is no claim. B. His claim is \_\_\_\_\_
- What is Bob's evidence? *The material of car A is not eco-friendly*  
A. There is no evidence. B. His evidence is (mark off use "\_\_\_\_\_")
- If so, which of the following do you think is true of Bob's evidence? (can choose more than one) *C*  
A. The evidence and claim are relevant.  
B. The evidence is sufficient.  
C. None of above
- What is Bob's reason? *eco-friendly materials are usually lighter*  
A. There is no reason. B. His reason is (mark off use "-----")
- If so, which of the following do you think is true of Bob's reason? (can choose more than one) *D*  
A. Reasons are related to evidence and claim  
B. Reasons are reasonable  
C. Reasons are comprehensive  
D. None of above
- Do you agree with Bob?  
A. yes B. no
- What evidence make you agree/not agree with Bob?
- Why the evidence make you agree/not agree with Bob?

**Scene 6: Now is the time for fellowship activities.**

Bob, Jane and other students who came to the University took part in the throwing contest. Whoever hits the most of yellow rubber ball wins. To make the game challenging, one person stands on point A on the platform and releases the ball while participant throws the purple ball at B.





They gathered the following information:

*Yellow ball is more massive than purple ball*

*Yellow ball is smaller than purple ball*

*Yellow ball is released without initial velocity*

*Air resistance has little effect on the balls' motion*

Bob: *It's more likely to hit the ball if you aim it under the ball. The yellow ball has more mass, so it falls faster and is more likely to hit if you aim down.*

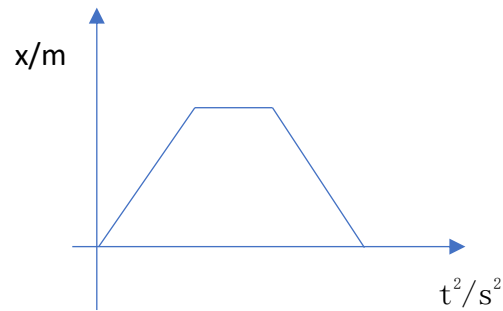
Jane: *I think it's more likely to hit the ball if aim directly at it.*

1. What is Bob's claim? *It's more likely to hit the ball if you aim it under the ball*  
A. There is no claim. B His claim is \_\_\_\_\_
2. What is Bob's evidence? *The yellow ball has more mass*  
A. There is no evidence. B. His evidence is (mark off use "\_\_\_\_\_")
3. If so, which of the following do you think is true of Bob's evidence? (can choose more than one) **C**  
A. The evidence and claim are relevant.  
B. The evidence is sufficient.  
C. None of above
4. What is Bob's reason? *The yellow ball has more mass, so it falls faster*  
A. There is no reason. B. His reason is (mark off use "-----")
5. If so, which of the following do you think is true of his reason? (can choose more than one) **D**  
A. Reasons are related to evidence and claim  
B. Reasons are reasonable  
C. Reasons are comprehensive  
D. None of above
6. Which sentence is rebuttal? **A**  
A. There is no rebuttal. B. The rebuttal is (mark off use "~~~~~")
7. If so, which of the following do you think is true of the rebuttal? (can choose more than one)  
A. Accurately pointed out the other party's mistake  
B. Rebuttal based on appropriate evidence  
C. Rebuttal is reasonable  
D. None of above
8. Who do you agree with more?  
A. Bob B. Jane C. My own opinion\_\_\_\_\_
9. What information above would you use to support Bob/Jane/yourself?
10. How does the information support Bob/Jane/yourself?
11. Why do you agree less with Jane/Bob/both?

**Scene 7: The study tour was almost over. At the last communication meeting, the teacher showed them a black box covering a moving object.**

*An electronic screen outside the black box is attached to a sensor that displays the object's displacement and weight data. Up and right are positive directions. They discussed the motion*

of the object in the black box.



Time	t <sub>1</sub>	t <sub>2</sub>	t <sub>3</sub>	t <sub>4</sub>	t <sub>5</sub>	t <sub>6</sub>
Weight data	4kg	6kg	2kg	4kg	2kg	6kg

Bob: The object accelerates to the right and then slows down. In the  $x-t^2$  image, the displacement is always positive, indicating that it is moving to the right all the time. According to  $x=v_0t+(at^2)/2$ , the displacement first increases indicating accelerating, and then decreases indicating decelerating.

B: I don't agree. There is a change in the weight of the object, which means it is accelerating upward then decelerating.

- What is Bob's claim? *the object accelerates to the right and then slows down*  
 A. There is no claim. B His claim is \_\_\_\_\_
- What is Bob's evidence? *the displacement is always positive; the displacement first increases then decreases*  
 A. There is no evidence. B. His evidence is (mark off use "\_\_\_\_\_")
- If so, which of the following do you think is true of Bob's evidence? (can choose more than one) **A**  
 A. The evidence and claim are relevant.  
 B. The evidence is sufficient.  
 C. None of above
- What is Bob's reason? *indicating that it is moving to the right all the time. According to  $x=v_0t+(at^2)/2$ , the displacement first increases indicating accelerating, and then decreases indicating decelerating*  
 A. There is no reason. B. His reason is (mark off use "-----")
- If so, which of the following do you think is true of his reason? (can choose more than one) **A**  
 A. reasons are related to evidence and claim  
 B. reasons are reasonable  
 C. reasons are comprehensive  
 D. None of above
- Which sentence is rebuttal? *I don't agree. There is a change in the weight of the object. It means it is accelerating upward then decelerating*  
 A. There is no rebuttal. B. The rebuttal is (mark off use "~~~~~")
- If so, which of the following do you think is true of the rebuttal? (can choose more than one) **B**

- A. Accurately pointed out the other's mistake
  - B. Rebuttal based on appropriate evidence
  - C. The rebuttal is reasonable
  - D. None of above
8. Who do you agree with more?  
A. Bob    B. Jane    C. My own opinion\_\_\_\_\_
9. What evidence would you use to support Bob/Jane/yourself?
10. How does the information support Bob/Jane/yourself?
11. Why do you agree less with Jane/Bob/both?

This is the end of the journey. Did you enjoy it?



## Appendix 18 Test version II-students

### Scientific Argumentation Competence Test

(Thank you very much for your participation! Your serious and independent answer is very important to this research. Sincerely, please answer seriously and independently)

Name: \_\_\_\_\_ (The researcher will keep it confidential for you)

**Do two items to warm up first! (Please note the time you started answering the question)**

The first two tasks are to help you understand what is **claim**, **evidence** (which can support a claim), **reason** (which can explain the connection between evidence and claim, that is, why certain evidence supports a certain claim), and **rebuttal** (point out the error of an argument and explain why), but not to show what a good argument is.

**1. Match in the blanks. Which ball has bigger inertia?**

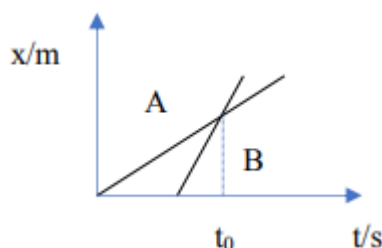
	Velocity	Mass
Ball A	5m/s	5kg
Ball B	2m/s	10kg

- 1) The velocity is high, and it is not easy to stop, so the inertia is large.
- 2) The velocity of ball A is 5m/s, which is greater than that of ball B.
- 3) Inertia is only related to mass, not velocity. It should be that the inertia of ball B is large.
- 4) The inertia of ball A is greater.

**Please select from 1-4 to fill in the brackets below:**

Claim is ( 4 ) ; Evidence is ( 2 ) ; Reason is ( 1 ) ; Rebuttal is ( 3 )

**2. Match in the blanks. Did A and B meet?**



- 1) The velocity of A is always less than that of B, so the two did not meet.
- 2) The intersection point represents the same displacement of A and B at the same time.
- 3) A met B once.
- 4) The two lines intersect at  $t_0$ .

**Please select from 1-4 to fill in the brackets below:**

Claim is ( 3 ) ; Evidence is ( 4 ) ; Reason is ( 2 ) ; Rebuttal is ( 1 )

Bob and Jane went to Y University for a study tour, and what they saw and heard sparked so much thinking and discussion about physics!

(Please answer based on what you have learned, and the information given in the question.

Multiple choice questions **have one or more options.**)

Formula scaffold:  $F=ma$ ;  $x=v_0t+at^2/2$ ;  $v^2-v_0^2=2ax$

**Scene 1: It started to rain on the way to the station. The two were anxious to catch the train, but the taxi driver drove slower.**



**Bob: I think water reduces friction. The car slows down in the rain, because the friction between the tire and the road surface becomes smaller, and it is not easy to brake.**

**Jane: I think water adds friction, like when you're counting money, it's easier to separate the money by dipping it in your finger. In rainy days, the vehicle becomes slow is not because of the smaller friction, but because the sight is not clear, affecting the driver to evaluate the road condition.**

1. What is Bob's claim? **water reduces friction**  
 A. There is no claim. B His claim is \_\_\_\_\_
2. What is Bob's evidence? **car slows down in the rain**  
 A. There is no evidence. B. His evidence is (mark off use "\_\_\_\_\_")
3. What is Bob's reason? **the friction between the tire and the road surface becomes smaller, and it is not easy to brake**  
 A. There is no reason. B. His reason is (mark off use "-----")
4. What is Jane's claim? **water adds friction**  
 A. There is no claim. B Her claim is \_\_\_\_\_
5. What is Jane's evidence? **when you're counting money, it's easier to separate the money by dipping it in your finger**  
 A. There is no evidence. B. Her evidence is (mark off use "\_\_\_\_\_")
6. What is Jane's reason? **A**  
 A. There is no reason. B. Her reason is (mark off use "-----")
7. Which sentence is rebuttal? **In rainy days, the vehicle becomes slow is not because of the smaller friction, but because the sight is not clear, affecting the driver to judge the road condition.**  
 A. There is no rebuttal. B. The rebuttal is (mark off use "~~~~~")

*Later, they got the following experimental data from internet.*

*Table 1 maximum static friction between shoes and the ground*

	Shoes	First (N)	Second (N)	Third (N)	Average (N)
Asphalt surface	Dry	2.84	2.84	2.84	2.84
	Wet	2.35	2.25	2.45	2.35
Cement floor	Dry	2.84	2.74	2.79	2.79
	Wet	2.06	1.86	1.86	1.93
Terrazzo floor	Dry	1.72	1.76	1.76	1.75
	Wet	1.86	1.91	1.91	1.90

*Table 2 maximum static friction between leather and paper currency*

	First (N)	Second (N)	Third (N)	Average (N)
Dry	0.93	0.93	0.93	0.93

One water spray	1.37	1.42	1.27	1.35
Two water spray	1.47	1.37	1.47	1.44
Three water spray	1.37	1.27	1.32	1.32

Table 3. Maximum static friction between sandpaper

	Water volume	Average value (N)		Water volume	Average value(N)
	Ordinary sandpaper	Dry		1.1	Sandpaper with waterproof spray
One spray		1.3	One spray	1.25	
Two sprays		1.35	Two sprays	1.16	
Three sprays		0.98	Three sprays	1.04	

**Bob:** The effect of water on friction is material dependent. In Table 1, the friction between the sole and the asphalt and concrete floors decreases after adding water, and the friction between the terrazzo floor increases after adding water, and the contact surface material is the only changing variable in this experiment.

**Jane:** I don't agree. The friction between the leather and banknotes in Table 2 and the sandpaper in Table 3 increases when a small amount of water is added, and then decreases, suggesting that the effect of water on friction has nothing to do with the material, but is related to the amount of water.

8. Bob's evidence is highlighted using "—", which of the following is true of the evidence? **A**

- A. The evidence and claim are relevant.
- B. The evidence is sufficient.
- C. None of the above

9. Bob's reason is highlighted using "----". Which of the following is true of the reason? **AB**

- A. The reason is reasonable
- B. The reason is comprehensive
- C. None of the above

10. rebuttal is classified using "~~~~", which of the following is true of the rebuttal? **B**

- A. Point out the other's mistake
- B. Rebuttal is evidence-based
- C. Rebuttal is reasonable
- D. None of the above

11. Who do you agree with more?

- A. Bob
- B. Jane
- C. My own opinion \_\_\_\_\_

12. What evidence makes you agree with Bob/Jane/Yourself?

13. How does the evidence support Bob/Jane/yourself?

14. Why do you think Jane/Bob/both is wrong?

**Scene 2: They finally catch the train before it leaves. There were several children in the carriage, and one of them flew his toy helicopter.**



**Bob: It's dangerous to play a helicopter in the carriage.**

**Jane: Why?**

**Bob: The helicopter will not continue to hover steadily after driving. It will bump into people or the door of the carriage.**

**Jane: No, neither we nor our bags slid back after driving. The movement of stuff in the carriage and the carriage are the same.**

1. What is Jane's claim? **It's not dangerous to play a helicopter in the carriage**  
 A. There is no claim. B Her claim is \_\_\_\_\_
2. What is Jane's evidence? **neither we nor our bags slid back after driving**  
 A. There is no evidence. B. Her evidence is (mark off use "\_\_\_\_\_")
3. What is Jane's reason? **The movement of stuff in the carriage and the carriage are the same**  
 A. There is no reason. B. Her reason is (mark off use "-----")
4. Which sentence is rebuttal? **No, neither we nor our bags slid back after driving. The movement of stuff in the carriage and the carriage are the same.**  
 A. There is no rebuttal. B. The rebuttal is (mark off use "~~~~~")
5. Who do you agree with more?  
 A. Bob B. Jane C. My own opinion\_\_\_\_\_
6. Which of the following facts/theories support Bob/Jane/yourself? **AC**  
 A. Bodies that are not subjected to external forces tend to remain in motion  
 B. When a train is traveling in a straight line at a constant speed, the object in the carriage is traveling at the same speed as the train  
 C. The train accelerates when it starts  
 D. When the train accelerates, the front carriage drives the rear carriage
7. How does the evidence support Bob/Jane/yourself?
8. Why do you think Jane/Bob/both is wrong?

**Scene 3: On their first day at University Y, the two visited the university's physics laboratory. The first activity they participated in was the hands-on measurement of the kinetic friction factor  $\mu$  they had just learned.**

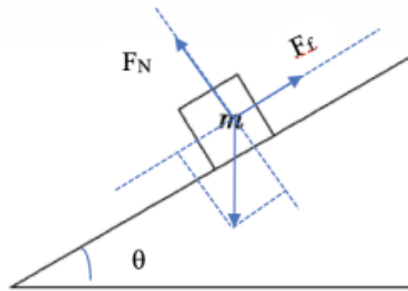
Bob finds the following equipment:

- A. Inclined surface with adjustable angle
- B. Wooden block
- C. Balance
- D. Protractor

**Bob: These devices can measure  $\mu$ . Adjust the inclined plane to let the block slide down, the mass of the block measured by the balance is  $m$ , and the inclination angle of the**

inclined plane measured by the protractor is  $\theta$ . Then, from  $F_f = \mu F_N$ ,  $\mu$  can be obtained.

Jane:  $F_f$  is unknown, and  $\mu$  cannot be derived.



1. Bob's evidence is highlighted using "—", which of the following is true of the evidence? **A**
  - A. The evidence and claim are relevant.
  - B. The evidence is sufficient.
  - C. None of the above
2. Bob's reason is highlighted using "----". Which of the following is true of the reason? **A**
  - A. The reason is reasonable
  - B. The reason is comprehensive
  - C. None of the above
3. Rebuttal is classified using "~~~~", which of the following is true of the rebuttal? **A**
  - A. Point out the other's mistake
  - B. Rebuttal is evidence-based
  - C. Rebuttal is reasonable
  - D. None of the above
4. Who do you agree with more?
  - A. Bob
  - B. Jane
  - C. My own opinion \_\_\_\_\_
5. Why do you think Jane/Bob/both is wrong?

**Scene 4: Later, they measured the mass of the car on an existing device in the laboratory and recorded the following information. Track A and B are the same.**

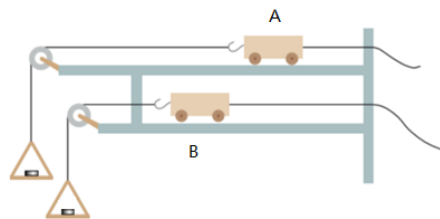



Table 3

Track	Initial speed (m/s)	Displacement (m)	Diameter of rope (m)	Mass of weight (kg)	Time of motion (s)	Car material
A	0	0.2	0.04	0.02	4	Non-Eco-friendly
B	0	0.2	0.02	0.02	2	Eco-friendly

**Bob: Car A has larger mass. The body of A is not an environmentally friendly material, and the environmentally friendly material should be lighter.**



**Jane: Car A indeed has larger mass, but the mass of the car has nothing to do with whether it is environmentally friendly or not.**

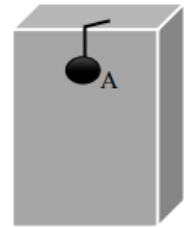
1. Rebuttal is classified using "", which of the following is true of the rebuttal? **A**
  - A. Point out the other's mistake
  - B. Rebuttal is evidence-based
  - C. Rebuttal is reasonable
  - D. None of the above
2. Who do you agree with more?
  - A. Bob
  - B. Jane
  - C. My own opinion\_\_\_\_\_
3. What evidence makes you agree with Bob/Jane/Yourself?
4. How does the evidence support Bob/Jane/yourself?

**Scene 5: Now is the time for leisure activities.**

The students organized a throwing competition, and the one who hit the most A balls is the winner. To make the game challenging, one person stands on a raised platform and releases a ball A while the participant throws a ball B on the ground.


*They gathered the following information:*

- a. Ball A is more massive than ball B
- b. Ball A is smaller than ball B
- c. Ball A is released without initial velocity
- d. Air resistance has little effect on the balls' motion



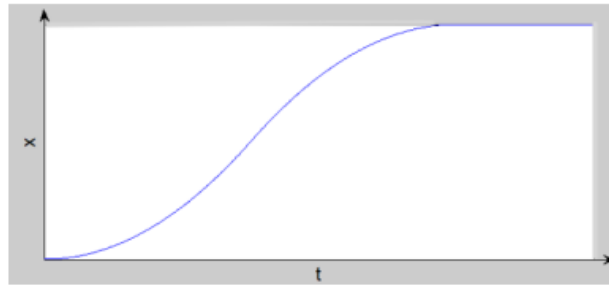
Bob: It's more likely to hit the ball if we aim it under the A ball. Ball A has larger mass, so it falls faster, and it is more likely to hit by aiming down.

Jane: I don't think so.

1. Bob's evidence is highlighted using "—", which of the following is true of the evidence? **C**
  - A. The evidence and claim are relevant.
  - B. The evidence is sufficient.
  - C. None of the above
2. Bob's reason is highlighted using "----". Which of the following is true of the reason? **C**
  - A. The reason is reasonable
  - B. The reason is comprehensive
  - C. None of the above
3. Which sentence is rebuttal? **A**
  - A. There is no rebuttal.
  - B. The rebuttal is (mark off use "")
4. Who do you agree with more?
  - A. Bob
  - B. Jane
  - C. My own opinion\_\_\_\_\_
5. What information above would you use to support Bob/Jane/yourself?
6. How does the information support Bob/Jane/yourself?
7. Why do you agree less with Jane/Bob/both?

**Scene 6: The study tour was almost over. At the last communication meeting, the teacher showed them a black box with a moving object in it.**

An electronic screen outside the black box is attached to a sensor that displays the object's displacement and weight data. Up and right are positive directions. They discussed the motion of the object in the black box.



<b>Time</b>	$t_1$	$t_2$	$t_3$	$t_4$
<b>Weight data</b>	4kg	6kg	2kg	4kg

**A: The displacement in the figure increases positively, the oblique first increases and then decreases, and the weight data first increases and then decreases. The slope of the x-t image represents velocity, and it is known that the right direction is positive, so the object moves to the right, first accelerating and then decelerating.**

**B: I don't agree. The movement of the object to the right cannot explain the change of the weight data. The change of the weight from large to small indicates that the object is moving up and down.**

1. Bob's evidence is highlighted using "—", which of the following is true of the evidence?

AB

- A. The evidence and claim are relevant.
- B. The evidence is sufficient.
- C. None of the above

2. Bob's reason is highlighted using "----". Which of the following is true of the reason? A

- A. The reason is reasonable
- B. The reason is comprehensive
- C. None of the above

3. Rebuttal is classified using "~~~~", which of the following is true of the rebuttal? ABC

- A. Point out the other's mistake
- B. Rebuttal is evidence-based
- C. Rebuttal is reasonable
- D. None of the above

4. Who do you agree with more?

- A. Bob
- B. Jane
- C. My own opinion\_\_\_\_\_

5. What evidence would you use to support Bob/Jane/yourself?

6. How does the information support Bob/Jane/yourself?

7. Why do you agree less with Jane/Bob/both?



**The journey ends here! Hope you are happy and rewarded!**

**Thanks for your hard work! Please answer the following questions truthfully.**

- How long did it take you to do this test? ( )

- How seriously did you do the test? ( )  
1. Very serious 2. Seriously 3. Not sure 4. Not serious 5. Very not serious
- Did you complete this test independently? ( )  
1. Completely independent 2. Partially consulted or sought help 3. Almost all consulted or sought help
- Do you find this set of questions difficult? ( )  
1. Very difficult 2. Relatively difficult 3. Not sure 4. Relatively easy 5. Very easy

## Appendix 19 Test version III

### Scientific Argumentation Competence Test

Please answer according to the information given in the test and the background knowledge you have learned.

School \_\_\_\_\_ Class \_\_\_\_\_ Name \_\_\_\_\_ Gender \_\_\_\_\_

#### Explanation of the four elements of scientific argument:

**Claim:** an opinion or conclusion on an issue.

**Evidence:** data or facts used to support a claim.

**Reason:** an explanation of the link between evidence and an opinion, i.e., why a certain piece of evidence supports a certain claim.

**Rebuttal:** questioning and weakening the arguments of others.

#### I Identification of argument elements—This is to see whether you can identify the four elements of argument in a piece of given argumentation

##### Problem 1: Which has greater inertia, ball A or ball B?

	Velocity	Mass
Ball A	5m/s	5kg
Ball B	2m/s	10kg

- 1) Inertia is an inherent property of objects, which is only related to mass. The greater the mass, the greater the inertia.
- 2) The mass of ball B is 10kg, which is larger than that of ball A.
- 3) The ball with high velocity is less easy to stop, and it is not necessarily only related to mass.
- 4) Ball B has greater inertia.

#### Please select from 1-4 to fill in the brackets below:

Claim is ( 4 ) ; Evidence is ( 2 ) ; Reason is ( 1 ) ; Rebuttal is ( 3 )

##### Problem 2: Who has a larger scale in hand?

Bob and Jane are about the same height, but Bob weighs 80 kg and Jane 60 kg. They each hold a weight scale in their hands to push each other, neither of them stepped back. Who has the bigger scale in his hand?



Li said: “① The scale in Jane’s hand indicates a large number.”

Jo said, “② Why do you think so?”

Li said: “③ Bob weighs 80 kg, heavier than Jane.”

Jo said: “④ What is the relationship between Bob's 80 kg and the scale of Jane?”

Li said: “⑤ Generally heavier people are more powerful, so the scale in Jane's hand will be pushed more.”

Jo said: “⑥ You’re wrong, heavy people are not necessarily powerful.”

Please select a serial number from ①-⑥ and fill in the following brackets (single choice):

Claim is ( 1 ), Evidence is ( 3 ), Reason is ( 5 ), Rebuttal is ( 6 )

II. Evaluate the elements of argumentation (single choice) - judge whether the elements of argumentation conform to the given indicators

Problem 3: Does water increase or decrease friction?

“I think water increases friction,” says Bob. “When counting money, it's easier to count with your fingers dipped in water.”

Jane said: “You are too one-sided, and the tires tend to slip when the road is wet.”



1. Bob believes that the single-underlined text is his evidence, which of the following do you think his evidence fits into? ( A ) (one or more choice)

- A. Bob’s evidence is relevant to his claim
- B. Bob’s evidence is sufficient to prove that his claim is right
- C. None of the above

2. Jane thinks that the double-underlined text is her rebuttal to Bob. Which of the following do you think her rebuttal fits into? ( ABD ) (one or more choice)

- A. Jane points out Bob’s deficiency
- B. Jane proves Bob’s deficiency with appropriate evidence
- C. Jane provides her own claim and explains about it
- D. All of what Jane says is right
- E. None of the above

Problem 4: How can they hit the black ball?

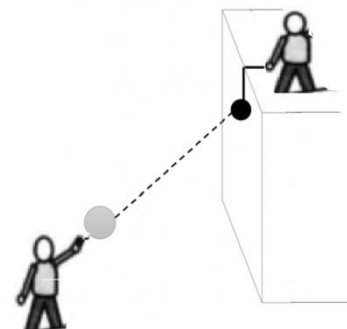
Four students are participating in a throwing competition, and the player who hits more black balls with gray balls wins. While the participants stand on the ground to throw the gray ball, one person releases the black ball from the high platform at the same time.

Some facts about the two balls:

- a. Black balls are heavier than gray balls
- b. Black balls are smaller than gray balls
- c. Black balls are released without initial velocity
- d. Air resistance has little effect on the balls’ motion

Some information about gray ball’s motion:

- e. If the gray ball is not subject to gravity, it will move in a straight line toward the direction it was thrown (dotted line in the figure).
- f. Gravity makes the vertical displacement of the gray ball  $gt^2/2$  lower than without gravity.



Bob: “Aiming below the black ball is easier to hit it. Because the black ball has larger mass.”

1. Bob’s evidence is marked using the single underline, which of the following is true of his evidence? ( C ) (one or more choice)

- A. Bob’s evidence is relevant to his claim
- B. Bob’s evidence is sufficient to prove that his claim is right

C. None of the above

**Jane: “I also think we should aim below the black ball. According to facts c and d, I feel that the black ball will fall faster, and the gray ball can only hit it by aiming down.”**

2. What is Jane’s evidence? ( facts c and d )

A. She provides no evidence B. Her evidence is: (please mark with single underline)

3. Jane’s reason is marked using the dotted line, which of the following is true of her reason?

( C ) (one or more choice)

A. Jane explains her evidence correctly

B. Jane provides sufficient reason to explain why her evidence supports her claim

C. None of the above

**Li: “I disagree with you two. I think we should aim directly at the black ball. According to c, d, e and f, the black ball is in free-fall and the gray ball is in projectile motion, we should decide based on their movement.”**

4. Li’s rebuttal against Bob and Jane is marked using the double underline, which of the following is true of his rebuttal? ( CD ) (one or more choice)

A. Li points out Bob and Jane’s deficiency

B. Li proves Bob and Jane’s deficiency using appropriate evidence

C. Li provides his own claim and explains about it

D. All of what Li says are right

E. None of the above

**III Production of argument—This is to see whether you can formulate your own argument when facing scientific issues**

**Problem 5: Will the toy helicopter suspended in the carriage collide?**



**Some facts about motion, trains, and toy helicopters:**

a. The helicopter remote control has three function buttons: ascend, descend, and keep hovering.

b. Objects that are not subject to external forces tend to maintain their original state of motion.

c. There will be acceleration when the train starts and during its running.

d. The train has a maximum speed limit.

**Bob: “I think there will be a collision, like hitting a person or something else in the cabin.”**

**Jane: “The luggage on the train does not slide, so the helicopter will hover stably after it rises, and there will not be a collision.”**

1. What is Bob’s claim?

A. He provides no claim B. His claim is there will be a collision

2. What is your claim in terms of the question Bon and Jane are discussing? Which fact(s) from a-d could be used as evidence to support yourself? abc

3. Why do the fact(s) you choose support your claim?

4. How would you rebut the side that you disagree with?

**Problem 6: How does water affect friction?****Experimental data about water and friction:***Table 1. Maximum static friction between leather and paper currency*

Water volume	Average friction (N)
Dry	0.93
One spray	1.35
Two sprays	1.44
Three sprays	1.32

*Table 2. Maximum static friction between sandpaper*

Ordinary sandpaper	Water volume	Average value (N)	Sandpaper with waterproof spray	Water volume	Average value(N)
	Dry	1.1		Dry	1.3
	One spray	1.3		One spray	1.25
	Two sprays	1.35		Two sprays	1.16
	Three sprays	0.98		Three sprays	1.04

**Some facts about water and friction:**

a. When water is in contact with a solid, and the attraction of solid molecules to water molecules is greater than the cohesive force of water molecules at the contact point, infiltration will occur.

b. The waterproof material utilizes the not-infiltration phenomenon of water.

c. Friction is always in the opposite direction of movement (or movement trend).

**Li: “Friction is affected by the volume of water. Table 1 can support my claim. When there is too little water, the surfaces of the two objects may still be in direct contact and are not covered with water, so the friction is increased.”**

**May: “I disagree with you.”**

1. What is your opinion in terms of how water affects friction? What is your evidence by saying this? (Evidence could be the knowledge you have learned, your observations from daily life or the data and facts provided in this task)

2. Why does your above-mentioned evidence support your claim?

3. If you were May, how would you rebut Li?

**Problem 7: Which fuel should be used?**

Bob and Jane opened a new factory, and they were considering whether to use petroleum or ethanol as the fuel for the factory. The information collected by them about the performance and price of the two fuels, global climate change research, and the impact of global climate change are as follows.

**Information 1: Performance and price of the two fuels**

Table 1 Comparison of two fuels on price and performance

Fuel	Energy produced (kJ/g)	CO <sub>2</sub> emission (mg/kJ)	Price (RMB/Ton)
Petroleum	43.6	78	2500
Ethanol	27.3	59	6000

**Information 2: The latest research of predicting global temperature**

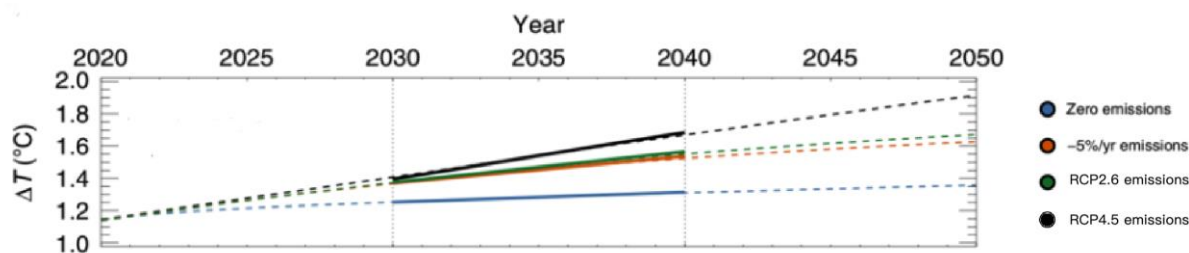


Figure 1 Prediction map of temperature change rate under four CO<sub>2</sub> emission conditions (the ordinate is the height of temperature rise, and the abscissa is time)

**Figure 1 description:** Zero emissions: CO<sub>2</sub> emissions keeps zero from 2020; -5%: CO<sub>2</sub> emissions reduced by 5% every year from 2020; RCP2.6: Start to reduce CO<sub>2</sub> emissions in 2020 and reduce to zero in 2100, and the temperature will rise below 2°C by 2100; RCP4.5: CO<sub>2</sub> emissions are greater than RCP2.6, and the temperature will rise by 2-3°C by 2100)

**Information 3: The latest research of predicting the survival of polar bears this century**

Table 2 Polar bear survival prediction

Global rising temperature (Celsius)	Polar bear survival situation
4	May lose almost all polar bears at the end of this century
2	May survive this century

1. Which fuel do you think is better for this newly opened factory?  
A. Petroleum      B. Ethanol
2. What evidence supports your decision?
3. Why do the evidence support your decision?
4. If someone chooses the other fuel, how would you rebut him/her?

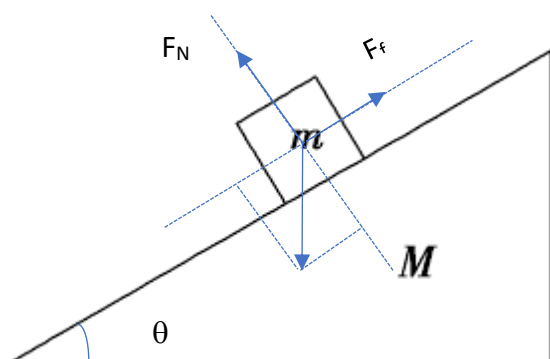
**IV Comprehensive tasks—This is to see whether you can identify and evaluate other’s arguments and generate your own arguments.**

**Problem 8: Whose experimental instruments can be used to measure  $\mu$ ?**

Bob and Jane find different sets of experimental instruments to measure the kinetic friction coefficient, and they all argue that their instruments can be used to measure  $\mu$ .

**Bob’s instruments: Angle adjustable bevel, wooden block, protractor**

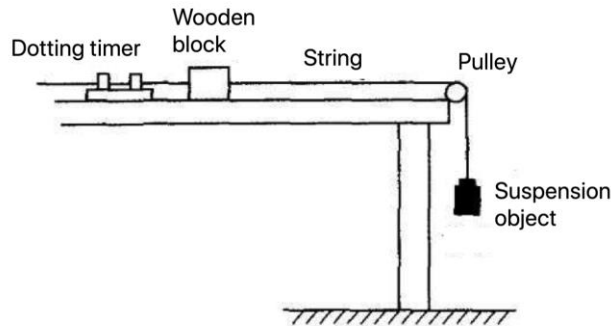
Bob: “These instruments can make the wooden block slide down at a uniform velocity. The force of the wood block is shown in the figure below, then  $F_f = \mu F_N$ ,  $F_f = mg \sin \theta$ ,  $F_N = mg \cos \theta$ , substitute the measured  $\theta$  into the formula to get  $\mu = \sin \theta / \cos \theta$ .”



**Jane’s instruments: Long plank with pulley, wooden block, dotting timer and paper tape, suspension object, string.**

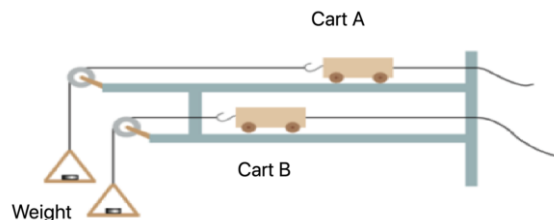


Jane: “These instruments can be installed into the device shown below. After the suspension object falls and hits the ground, the block is only subjected to kinetic friction force to do a uniform deceleration movement (acceleration is  $a$ ), during which time  $F_f = \mu mg = ma$ , the unknown quantity  $a$  can be obtained by analyzing the dots on the paper tape, so  $\mu = a/g$  can be obtained.”



- Which part of the above argument is a rebuttal to the other person? ( A )  
 A. There is no rebuttal      B. Rebuttal is (please mark with wavy lines)
- Jane’s reason is marked using the dashed line, which of the following is true of her reason? ( AB ) (one or more choice)  
 A. Jane explains her experimental device correctly  
 B. Jane provides sufficient reason to explain why this device can measure  $\mu$   
 C. None of the above
- Whose instruments do you think can measure  $\mu$ ?  
 A. Bob    B. Jane    C. None of them
- How would you rebut against the one(s) whose instrument cannot measure  $\mu$ ?

**Problem 9: Which cart has bigger mass?**



**Facts about tracks and carts:**

- Cart A is solid, cart B is hollow
- Both carts have wooden wheels
- The track of A is wooden, and the track of B is metal
- Car A is made of ordinary materials, and car B is made of environmentally friendly materials
- The mass of the weight connected on the two carts are the same

**The data of the two carts moving for 2 seconds under the pull of the weight:**

Cart	Initial velocity $V_0$ (m/s)	Displacement $x$ (m)	Movement time (s)
A	0	0.4	2
B	0	0.6	2

**The friction coefficients of some surfaces:**

Material	wood-ice	wood-metal	steel-ice	wood-wood
Friction coefficient	0.03	0.20	0.02	0.30

**Bob:** “Cart A has bigger mass. It can be judged from fact d. Car B is an environmentally friendly material and should use less material, so it is lighter.”

**Jane:** “Environmental friendliness has nothing to do with materials and weight, and environmentally friendly materials do not necessarily mean small mass. It is true that the mass of car A is larger, but it should be judged by fact a, because the mass of a solid car must be larger than a hollow one.”

1. What is Bob’s reason? ( ) **Car B is an environmentally friendly material and should use less material, so it is lighter**

A. He doesn’t provide reason B. His reason is: (please mark using dashed line)

2. Bob believes his evidence is fact d. Which of the following is true of his evidence? ( C ) (one or more choice)

A. Bob’s evidence is relevant to his claim

B. Bob’s evidence is sufficient to prove that his claim is right

C. None of the above

3. Jane’s rebuttal is marked using the single underline, which of the following is true of her rebuttal? ( AC ) (one or more choice)

A. Jane points out Bob’s deficiency

B. Jane proves Bob’s deficiency using appropriate evidence

C. Jane provides her own claim and explains it

D. All of what Jane says are right

E. None of the above

**Li:** “I think cart B has bigger mass. According to facts b, c, e and the first table. They move for the same time and the displacement of the B is large, indicating that B has a large acceleration. The mass of the weight is the same, so the pulling force of the two carts is the same. Therefore, cart B has bigger mass.”

4. Li believes that the dashed line marks his reason. Which of the following is true of his reason? ( A ) (one or more choice)

A. Jane explains her evidence correctly

B. Jane provides sufficient reason to explain why her evidence supports her claim

C. None of the above

5. What is your opinion on the issues discussed by Bob, Jane, and Li? What evidence would you use to support yourself?

6. Why does the evidence support your claim?



**Thank you for participating in this study! There are still a few small questions, please answer them truthfully.**

• The researcher would like to invite you to an interview around 30-minute to chat about your experience of taking the test and learning science. Would you like to participate? If Yes, please leave your contact \_\_\_\_\_ (WeChat number, QQ number or mobile phone number can be used).

• How long did it take you to do this test? ( )

• How seriously did you do the test? ( )

1. Very serious 2. Seriously 3. Not sure 4. Not serious 5. Very not serious

- Did you complete this test independently? ( )
  1. Completely independent
  2. Partially consulted or sought help
  3. Almost all consulted or sought help
- Do you find this set of questions difficult? ( )
  1. Very difficult
  2. Relatively difficult
  3. Not sure
  4. Relatively easy
  5. Very easy

## Appendix 20 Test specification

- General description of the test

This test aims to assess Chinese high school students' scientific argumentation competence from their ability of identifying SA elements, evaluating SA elements, and producing SA elements. In accordance with our framing of each component, multiple-choice items, open-ended items, and selection items are included in the test. Considering Chinese students' familiarity with paper-pencil test, the test adopts a paper-pencil format. Due to the constraint of the test format, this test does not aim to assess the social aspect of SA and capture the dynamic process of students' engagement in SA.

The test aims at minimizing the needed content knowledge. However, given the impossibility of rigorously control the cognitive demand of each task except for that needed by argumentation, the test aims to pay attention to the content knowledge, involved information, and topic familiarity to decide the complexity of each task.

Considering high school students' limited time of participating in the assessment, the test aims to be designed as can be finished around 45 mins.

- Item/task summary

Task	Item	SA element	Context	Task complexity	Score
1	1	Identification	Science	Less (Less information; More familiar; Less content knowledge)	1
2	2.1	Evaluate evidence	Science	Less (Less information; More familiar; Less content knowledge)	1
	2.2	Evaluate rebuttal			1
3	3.1	Evaluate evidence	Science	More (More information; More familiar; More content knowledge)	1
	3.2	Identify evidence			1
	3.3	Evaluate reason			1
	3.4	Evaluate rebuttal			1

4	4.1	Produce evidence	Science	Less (Less information; More familiar; More content knowledge)	2
	4.2	Produce reason			3
	4.3	Produce rebuttal			2
5	5.1	Produce evidence	Science	More (More information; Less familiar; Less content knowledge)	2
	5.2	Produce reason			3
	5.3	Produce rebuttal			2
6	6.1	Produce evidence	Social science		2
	6.2	Produce reason			3
	6.3	Produce rebuttal			3
7	7.1	Identify reason	Science	More (More information; More familiar; More content knowledge)	1
	7.2	Evaluate rebuttal			1
	7.3	Evaluate evidence			1
	7.4	Evaluate reason			1
	7.5	Produce evidence			2
	7.6	Produce reason			3

- Test grid

SAC component	SAC element	Number of items	Item format
Identification of SA	Claim	/	
	Evidence	1	Selection
	Reason	1	Selection
	Rebuttal	/	
	All elements	1	Match
Evaluation of SA	Evidence	3	Multiple-choice question
	Reason	2	Multiple-choice question
	Rebuttal	3	Multiple-choice question
Production of SA	Evidence	4	Open-ended question
	Reason	4	Open-ended question
	Rebuttal	3	Open-ended question

## Appendix 21 Test version IV

### Scientific Argumentation Competence Test

Thank you for your participation! Please answer independently.

Class \_\_\_\_\_ Name \_\_\_\_\_ Gender \_\_\_\_\_ Birthdate \_\_\_\_\_

#### Explanation of the four elements of scientific argument:

**Claim:** an opinion or conclusion about an issue.

**Evidence:** data or facts used to support a claim.

**Reason:** an explanation of the connection between evidence and claim, i.e., why a certain piece of evidence supports a certain claim.

**Rebuttal:** Based on listening to and thinking about the arguments of others, analyse the disagreements, and use evidence to weaken the arguments of others.

### I. Identify the elements of an argument—identify the four elements of a given argument

#### Problem 1: Whose weight scale shows a larger number?

Bob and Jane are about the same height, but Bob weighs 80 kg and Jane 60 kg. They each hold a weight scale in their hands to push each other, neither of them stepped back. Whose scale shows a larger number?



Li said: “① The scale in Jane’s hand shows a larger number.”

Jo said, “② Why do you think so?”

Li said: “③ Bob weighs 80 kg, heavier than Jane.”

Jo said: “④ What is the relationship between Bob’s 80 kg and the scale in Jane’s hand?”

Li said: “⑤ Generally heavier people have more strength, so the scale in Jane’s hand will be pushed more.”

Jo said: “⑥ You are wrong, heavy people are not necessarily strong.”

Please select a serial number from ①-⑥ and fill in the following brackets (single choice):

Claim is ( 1 ), Evidence is ( 3 ), Reason is ( 5 ), Rebuttal is ( 6 )

### II. Evaluate the elements of an argument (single choice) - judge whether the argumentation elements meet the given criteria

#### Problem 2: Does water increase or decrease friction?

“I think water increases friction,” says Bob. “When counting money, it's easier to count with your fingers dipped in water.”

Jane said: “You are too one-sided, and the tires tend to slip when the road is wet.”



1. Bob believes that the single-underlined text is his evidence, which of the following do you think his evidence fits into? ( B )

- A. The evidence is irrelevant with his claim and cannot support the claim
- B. The evidence is relevant to his claim, showing that his claim may be right, but not sufficient
- C. There is sufficient evidence to establish that his claim is right

2. Jane thinks that the double-underlined text is her rebuttal to Bob. Which of the following do you think her rebuttal fits into? ( C )

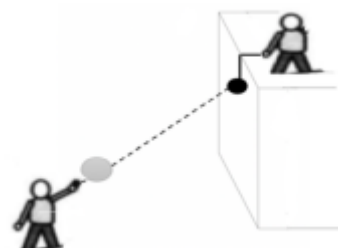
- A. Jane provides her claim without paying attention to Bob's argument
- B. Jane analyses Bob's argument without weakening his argument
- C. Jane weakens Bob's argument

### Problem 3: Where to aim to hit the black ball?

Three students are participating in a throwing competition, and the player who hits more black balls with grey balls wins. When the participants stand on the ground to throw the grey ball, one person releases the black ball from the high platform at the same time.

#### Some facts about the two balls:

- a. Black balls are heavier than the grey ball
- b. Black balls are released without initial velocity
- c. Air resistance has little effect on the balls' motion
- d. If the grey ball is not subject to gravity, it will move in a uniform straight line in the direction of throwing (dotted line)
- e. Gravity causes the vertical displacement of the grey ball to decrease by  $gt^2/2$  compared to when without gravity



**Bob said: "Aiming below the black ball is easier to hit. Because the black ball has larger mass."**

1. Bob believes that the single-underlined text is his evidence, which of the following do you think his evidence fits into? ( A )

- A. The evidence is irrelevant with his claim and cannot support the claim
- B. The evidence is relevant to his claim, showing that his claim may be right, but not sufficient
- C. There is sufficient evidence to establish that his claim is right

**Jane said: "I also think we should aim below the black ball. According to facts b and c, I have the feeling that the black ball will fall faster, and the grey ball can only hit it by aiming down."**

2. What is Jane's evidence? ( facts b and c )

A. She provides no evidence B. Her evidence is: (please mark it with single underline)

3. Jane thinks the dashed line is her reason, which of the following do you think her evidence fits into? ( A )

A. There is no connection between the reason and her evidence

- B. The reasons explain her evidence correctly, but cannot suggest that the evidence proves her claim being right
- C. The reason thoroughly illustrates the connection between her evidence and claim

**Li said: “I disagree with you two. I think we should aim directly at the black ball. Facts b, c, d, and e can support my claim.”**

4. Li thinks that the double-underlined text is his rebuttal. Which of the following do you think his rebuttal fits into? ( A )
- A. Li provides his claim without paying attention to Bob and Jane’s argument
  - B. Li analyses Bob and Jane’s argument without weakening their argument
  - C. Li weakens Bob and Jane’s argument

### III Generate an argument-give your own argument to a scientific problem

#### Problem 4: Will the toy helicopter hovering in the carriage collide?



#### Some facts about trains and the helicopter:

- a. The helicopter can only ascend, descend, and keep hovering
- b. The mass of the helicopter is 2 kg
- c. The train has accelerations when it starts and during its running
- d. The train has a maximum speed limit

**Bob said: “I think it will collide, such as hitting a person or other things in the cabin.”**

**Jane said: “The luggage in the cabin does not slide, so the helicopter will hover stably after it rises, and there will be no collision.”**

1. In terms of the question Bob and Jane are discussing, what is your claim? Which fact(s) from a-d can be used as evidence to support your claim?
2. Why do these fact(s) you choose support your claim?
3. How would you rebut the side(s) (Bob or/and Jane) who disagree(s) with you?

#### Problem 5: How does water affect friction?

#### Experimental data on water and friction:

Table 1. Maximum static friction between leather and banknotes

Addition of water	Average friction force (N)
Dry	0.93
One spray of water	1.35
Two sprays of water	1.44
Three sprays of water	1.32



Table 2. Maximum Static Friction Between Sandpapers

Normal sandpaper	Addition of water	Average friction force (N)	Sandpaper with waterproofing spray	Addition of water	Average friction force (N)
	Dry	1.1		Dry	1.3
	One spray of water	1.3		One spray of water	1.25
	Two sprays of water	1.35		Two sprays of water	1.16
	Three sprays of water	0.98		Three sprays of water	1.04

**Some facts about water and friction:**

- When water is in contact with a solid, the attraction of the solid molecules to the water molecules at the contact is greater than the attraction between the water molecules, the water droplets tend to adhere to solids, a phenomenon known as water infiltration.
- The waterproof material is based on a phenomenon that water does not have infiltration.
- Friction is always in the opposite direction of motion (or relative motion).

**Bob said: “I think as the amount of water increases, the friction increases first and then decreases. My evidence is the data in Table 1. The friction increases after one spray and two sprays and decreases after three sprays.”**

**Jane said: “I don’t agree with your statement.”**

- How do you think water affects the value of friction? What evidence would you use to support your claim?
- Why do the evidence(s) support your claim?
- How do you think Jane could rebut Bob’s argument?

**Problem 6: Which fuel should be used?**

The government of a middle-level economic region in eastern China plans to support several factories, either fossil fuels (petroleum) or biofuels (ethanol) can be chosen for the factories.

Table 3. Properties and prices of the two fuels

Fuel	Energy produced (kJ/g)	CO <sub>2</sub> released (mg/kJ)	Price (yuan/ton)
Petroleum	43.6	78	2500
Ethanol	27.3	59	4000

Table 4. Production and characteristics of the two fuels

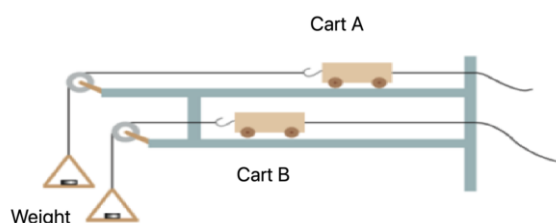
Fuel	Resource Regeneration	Production material	Storage and Transportation
Petroleum	Non-renewable	\	Easy
Ethanol	Renewable	Cereal crops	Difficulty

- Which fuel would you recommend? Which evidence(s) would you use to support your decision?

- Why do the evidence(s) support your decision?
- If someone chose a different fuel than you, how would you rebut his decision?

**IV Comprehensive task—This is to see whether you can identify and evaluate other’s arguments and generate your own arguments.**

**Problem 7: Which cart has larger mass?**



**Some facts about tracks and carts:**

- Cart A is solid, Cart B is hollow
- Both carts have wooden wheels
- The track of Cart A is wooden, and the track of Cart B is metal
- The mass of the weight connected on the two carts are the same

**The data of the two carts moving for 2 seconds under the pull of the weight:**

Table 5

Cart	Initial velocity $V_0$ (m/s)	Displacement $x$ (m)	Movement time (s)
A	0	0.4	2
B	0	0.6	2

**The friction coefficients of some surfaces:**

Table 6

Material	wood-ice	wood-metal	steel-ice	wood-wood
Friction coefficient	0.03	0.20	0.02	0.30

**Bob: “Cart A has bigger mass. It can be judged from fact a. Cart A is solid, it should use more material than the hollow cart, so it is heavier.”**

**Jane: “I don’t agree with you. Although Cart A is solid, it does not necessarily use more materials than Cart B.”**

- What is Bob’s reason? (Cart A is solid, it should use more material than the hollow cart, so it is heavier)
  - He doesn’t provide reason
  - His reason is: (please mark using dotted line)
- Jane believes that the double-underlined text marks her rebuttal. Which of the following is true of her rebuttal? ( B )
  - Jane provides her own claim without paying attention to Bob’s argument.
  - Jane analyzes Bob’s argument without weakening it
  - Jane weakens Bob’s argument

**Li: “I think Cart B has bigger mass. According to facts b, c, d and the two tables. They move for the same time and the displacement of the Cart B is large, indicating that Cart B has a large acceleration. The mass of the weight is the same, so the pulling force of the two carts is the same. The friction coefficient between Cart A and the track is larger.”**

**Therefore, Cart B has larger mass.”**

3. Li believes that the single underline marks his evidence. Which of the following is true of his evidence? ( C )
- A. The evidence is irrelevant with his claim and cannot support the claim
  - B. The evidence is relevant to his claim, showing that his claim may be right, but not sufficient
  - C. There is sufficient evidence to establish that his claim is right
4. Li believes that the dashed line marks his reason. Which of the following is true of his reason? ( B )
- A. There is no connection between the reason and his evidence.
  - B. The reason explains his evidence correctly but cannot suggest that the evidence proves his claim being right.
  - C. The reason thoroughly illustrates the connection between his claim and evidence
5. What is your opinion on the issues discussed by Bob, Jane, and Li? What evidence would you use to support yourself?
6. Why do the evidence(s) support your claim?



**Thank you for participating in this study! There are still a few small questions, please answer them truthfully.**

- The researcher would like to invite you to a 30-minute interview to chat about your experience of taking the test and learning science. Would you like to participate? If Yes, please leave your contact \_\_\_\_\_ (WeChat number, QQ number or mobile phone number can be used).
- How long did it take you to do this test? ( )
- Do you think you were given enough time to do the test? ( )  
1. Sufficient 2. Insufficient
- How seriously did you do the test? ( )  
1. Very serious 2. Seriously 3. Not sure 4. Not serious 5. Very not serious
- Did you complete this test independently? ( )  
1. Completely independent 2. Partially consulted or sought help 3. Almost all consulted or sought help
- Do you find this set of questions difficult? ( )  
1. Very difficult 2. Relatively difficult 3. Not sure 4. Relatively easy 5. Very easy

## Appendix 22 Scoring rubrics III

P-SA-Evidence	
Score	Descriptions
0	Irrelevant information
1	Partial evidence
2	Complete evidence

P-SA-Reason	
Score	Descriptions
0	Irrelevant information
1	Confusing logic/repeating evidence without trying to connect evidence and claim.
2	Connecting between claim and evidence but with minor incorrect knowledge understanding or incoherent logic.
3	Accurate and coherent reason connecting claim and evidence successfully respectively.

P-SA-Rebuttal	
Score	Description
0	Vague or irrelevant information
1	Emphasizing their own argument without analysing others.
2	Attending to others' argument without a comprehensive and coherent explanation of why it is not appropriate.
3	Weakening others' argument with evidence and coherent explanation respectively.

