The dynamics of reading development in L2 English for academic purposes

Gui, Min; Chen, Xiaokan; Verspoor, Marjolijn

Published in:
System

DOI:
10.1016/j.system.2021.102546

Publication date:
2021

System 100 (2021) 102546

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

The dynamics of reading development in L2 English for academic purposes

Min Gui [a], Xiaokan Chen [a], Marjolijn Verspoor [b,c,*]

[a] Research Institute of Foreign Languages, School of Foreign Languages and Literature, Wuhan University, China
[b] Department of English Language and Culture, University of Groningen, Groningen, Netherlands
[c] University of Pannonia, Veszprem, Hungary

ARTICLE INFO

Keywords:
L2 reading development
English for academic purposes
Variability
Strategies
CDST

ABSTRACT

In a mixed-methods approach, this study investigates the complex and dynamic developmental trajectories of 27 Chinese Chemistry major undergraduates' English academic reading ability. Twelve parallel tests were designed, validated, and used weekly during one semester. The analyses included a group pre-post design to measure academic reading gains, a regression analysis to predict beginning reading score with English proficiency and Chemistry knowledge as predictors, individual longitudinal case studies to measure variability and phase shifts, and a cluster analysis to discover (un)common developmental patterns. Finally, a qualitative study used interviews to discover difficulties in reading and strategies to overcome them. English proficiency predicted the initial reading score and the group gained significantly in academic reading. Each learner showed different non-linear patterns, and a cluster analysis revealed few similar patterns among learners. The high gainers showed relatively more variability over time and used more and a wider variety and more sophisticated learning and reading strategies to improve.

1. Introduction

The need to develop L2 academic reading ability has been reported by many studies (e.g., Anderson, 2015, pp. 95–109; Evans & Green, 2007; Hartshorn, Evans, Egbert, & Johnson, 2017; Jackson, 2005; Kaewpet, 2009; Pritchard & Nasr, 2004; Ward, 2001; Weir, Hawkey, Green, Unaldi, & Devi, 2012, pp. 37–119) and has been investigated in two main lines of research. The first line of research involves instructional approaches to develop learners' L2 academic reading ability (e.g., Amer, 1994; Carrell & Carson, 1997; Cheng, 2008; Kasper, 1995; Kuzborska, 2011; Martínez, 2002; Pritchard & Nasr, 2004). For example, in a pre-post design Pritchard and Nasr (2004) showed that students in the experimental group who used authentic materials and were asked to negotiate for meaning to understand difficult concepts gained significantly more than the control group who used simplified materials that were traditionally used at the institution. The second line of research examines factors influencing the development of L2 academic reading ability (e.g., Chen & Donin, 1997; Nergis, 2013; Usó-Juan, 2006). Nergis (2016) found that depth of vocabulary knowledge was not as strong a predictor of academic reading comprehension as English syntactic awareness and that metacognitive reading strategies have much to contribute to academic reading comprehension. Usó-Juan (2006) found that both domain knowledge and English proficiency contribute to EAP reading performance, but that English proficiency predicts EAP reading two to three times more than domain knowledge. Similar results were reported by Davis, Huang, & Yi, 2017. These two lines of research have employed cross-sectional

* Corresponding author. Department of English Language and Culture, University of Groningen, 5, 9712 CP Groningen, Netherlands.
E-mail addresses: guimin@whu.edu.cn (M. Gui), chenxk319@163.com (X. Chen), m.h.verspoor@rug.nl (M. Verspoor).

https://doi.org/10.1016/j.system.2021.102546
Received 23 July 2020; Received in revised form 1 April 2021; Accepted 11 May 2021
Available online 10 June 2021

approaches, with pre- and post-test designs, to examine learning outcomes in group studies. Moreover, both lines of research have tended to treat the influence of various factors in a static and linear approach, assuming they affect the whole process of academic reading development in the same way or influence all learners in the same manner. However, recently Cai and Kunnan (2019; 2020) have found that the contribution of some factors, such as reading strategy and background knowledge, fluctuate with the change in language proficiency. Thus, new perspectives, featuring dynamics and non-linearity of such relationships, are called for.

In the current study, we argue that learning to read for academic purposes in an L2 can be considered a complex process and taking Complex Dynamic System Theory (CDST) perspective can help in discovering how such academic reading ability develops. L2 reading in a subject area such as chemistry, can be argued to be a complex process as it is based on the interaction of many heterogeneous and interdependent components, including the reader's linguistic ability, personal characteristics, affective factors such as engagement and motivation, reading strategies, and domain knowledge (Carr & Levy, 1990; Douglas, 2000; Jordan, 1997). The learner's environment is also complex, with its textual, institutional, socio-cultural contexts interacting over time. Thus, academic reading ability is dynamic and adaptive as its internal components interact over time with textual contexts, reading goals, and the educational environment. Moreover, changes in one component (e.g., the goals or strategies of the reader) or the environment (e.g., reading different texts) may lead to changes in the whole system and this change is likely to be non-linear.

The current study is inspired by CDST and in particular by Lowie and Verspoor (2019), who suggest that longitudinal individual case studies and pre-post design group studies should complement each other. Group studies provide information about the relative weight of factors influencing L2 development and guide the identification of general factors influencing development, but because individuals have individual experiences and histories, they will behave idiosyncratically and the generalized factors found in group studies will not apply to all individuals in the same way. At the same time, what applies to one individual will not apply to the group. This is called the ergodicity problem.

So far, most CDST inspired studies have dealt with productive data, and in SLA specifically writing and oral production at different academic levels. To our knowledge, a receptive skill such as L2 academic reading ability has not been investigated from a CDST point of view.

Therefore, to add to the realm of research in both academic reading in an L2 and CDST, we focus on reading gains by means of 12 calibrated reading tests designed and validated for this study. In the group we attempt to find general predictors of initial reading scores, and in individuals we trace the process of academic reading development to see how they change over time. To see if indeed all individuals within the group behave differently, or if there may be sub-groups within the whole group, the current study also incorporates a cluster analysis. In addition, interview data—including types of strategies used to improve academic reading ability—will be examined to explore what may have motivated their changing behavior.

2. A CDST perspective on development and academic reading ability

Within CDST it is assumed that individuals will create their learning trajectories and will vary not only in their own behavior (referred to as *variability*) but also vary very much from each other (referred to as *variation*). The power of CDST in explaining L2 development has been evidenced by a series of empirical studies on L2 writing and oral development (e.g., Baba & Nitta, 2014; Larsen-Freeman, 2006; Lenzing, 2015; Lowie & Verspoor, 2019; Polat & Kim, 2014; Spoelman & Verspoor, 2010; Zheng, 2016). The findings are in general that no two learners are the same and that each learner has his or her own developmental path. One of the strongest claims in CDST is that variability in behavior is needed to progress. Thus Ortega (2011) argued that if variability is interpreted as a precursor for some important change in the system, we need analytical methods that are not only quantitative, as in the traditional perspective, but also innovatively different "because they are stochastic and non-causal, that is, based on probabilistic estimations that include the possibility of random variations and fluctuations tracked empirically over time (p 178)."

The idea that variability is needed to progress was evidenced in Lowie and Verspoor (2019), who attempted to find common patterns among small sub-groups of young high school learners, but concluded that no two learners were alike, but they did find that learners who showed relatively more variability were the ones who had progressed the most in L2 writing. In a recent replication of that study, Huang, Steinkraus and Verspoor (forthcoming) found the same. Of 22 Chinese university majors of English, who were traced on their L2 writing development over one year, the degree of variability rather than motivation, aptitude or working memory was a strong and significant predictor of L2 writing gains.

Variability is not a personality trait but just a symptom of behavioral choices as pointed out by Siegler (2006) in the field of developmental psychology. In his summary of twenty studies investigating children's learning in micro-genetic and longitudinal approaches, he concludes that especially early on in development, the learner discovers new approaches or strategies, and that when the learner uses them, the strategies are generally used inconsistently. Thus, at early stages of development one can expect relatively more variability. Secondly, he argues that learning reflects the addition of new strategies, with greater reliance over time on relatively advanced strategies, improved choices among strategies, and improved execution of strategies. The choices of strategies are less random and would therefore result in relatively less variability. Finally, although there is variability in the process of learning, learning tends to progress through a rather regular sequence of stages. In other words, even though learners tend to follow their own idiosyncratic developmental paths—causing variability—most will improve by using more advanced strategies more consistently, resulting in relatively less variability in each learner and concomitantly less variation among learners.

Siegler's views on strategies, albeit conceptualized in learning from a developmental view, seem to be in line with second language developmental research on the use of learning strategies (see Oxford, 2017 for a congruent overview). Successful learners generally use a greater number and a wider variety of learning strategies (McDonough, 1999).

To summarize, it is argued that L2 academic reading is influenced by both domain knowledge and English Proficiency, that

development is a dynamic process in which learners show variability in the process, with more ups and downs related to more progress, and variation in that no two learners will show the exact same development. Finally, successful learners will use a greater number and a wider variety of learning strategies.

The current study will be guided by five research questions.

1. To what extent does English proficiency or chemistry knowledge affect the initial academic reading score?

The hypothesis is that English proficiency will affect the initial academic reading score more than Chemistry knowledge, which would be in line with Usó-Juan (2006) and Davis, Huang, & Yi, 2017.

2.Does the group gain in academic reading scores after one semester?

The hypothesis is that through the L2 academic reading course itself and taking the frequent tests with feedback during the interviews, the group will gain significantly in reading scores. This would be in line with Pritchard and Nasr (2004) who showed that authentic materials and negotiating for meaning to understand difficult concepts was an effective way to increase academic reading ability.

3. To what extent are variability and/or phase shifts (sudden and radical changes) in academic reading scores related to gains in reading scores?

Variability concerns ups and downs over time, while phase shifts focus on sudden changes at specific points. The hypothesis is that higher degrees of variability (sometimes including phase shifts) are related to progress, operationalized as gains in reading scores.

4. Are there groups of individuals who share similar developmental trajectories?

In CDST it is argued that no two learners are exactly alike in every step of their developmental trajectory, but to test whether some sub-groups may be regarded "ergodic ensembles" we will examine if there are sub-groups of learners that are similar on all features, which in our study would be English Proficiency, Chemistry knowledge, degrees of variability, and gains. Based on Lowie and Verspoor (2019), the hypothesis is that few to no sub-groups are the same in all respects.

5. What factors do students comment on in terms of difficulties and strategies, and how can they be related to academic reading score gains?

Based on Siegler (2006) and L2 strategy research (Habók and Magyar, 2018; McDonough, 1999; Oxford, 2017), we expect successful learners to use a greater number and a wider variety of learning strategies.

3. Method

In this section we will first discuss how the reading test was designed and calibrated. Then we will discuss the actual experiment.

3.1. Test design

The design of the tests was guided by the test development procedures outlined by Bachman and Palmer (1996) and the principle of testing languages for specific purposes (Davies, 2001; Douglas, 2001). First, based on the analysis of the course objectives, we defined the construct "academic reading ability in chemistry" in the present study as "Chinese chemistry major undergraduates' ability to comprehend college level chemistry textbooks written in English" based on the objectives of the course. Second, with reference to reading abilities assessed in other studies (e.g., Alderson, Brunfaut, & Harding, 2015; Harding, Alderson, & Brunfaut, 2015; Jang, 2005), we described six types of academic reading operations, which served as the basis for writing testing items.

(1) To understand the meaning of basic terminology in chemistry.
(2) To identify and locate specific information in expository and instructive passages.
(3) To understand the information conveyed by graphs and tables.
(4) To understand sequences and main idea of instructive passages, such as requirements and sequences of laboratory operations.
(5) To understand the relationship of clauses and main content, such as the explanation of laws and principles, phenomena, causes and effects.
(6) To summarize the central idea of the text.

Third, 30 passages (24 expository, 6 instructive) were selected from five chemistry textbooks in English. Among them, 5 passages included graphs, and 3 had formulas. The average length of the passages is 240 words ($SD = 53.09$). The textbooks were *Fundamentals of General, Organic, and Biological Chemistry* (McMurry, Ballantine, Hoeger, Peterson, & Castellion, 2014), *The Extraordinary Chemistry of Ordinary Things* (Snyder, 2003), *Understanding Chemistry* (Lister & Renshaw, 2000), *Principles of General Chemistry* (Silberberg, 2013),

and *Advanced Chemistry through Diagrams* (Lewis, 2001).

Fourth, testing items were written by the authors and two professors in chemistry. Question types include vocabulary, True or False judgment, syntactic parsing, translation from English into Chinese, and summary. Finally, we paired these 30 passages with the testing items into 15 reading tests, based on the topic, length, and readability of passages indicated by Flesch-Kincaid Grade Level (Fraser, 2007). The information of topic, length, and readability of the 12 reading tests is presented in Table 1. The total score of each reading test was 100 points.

3.2. Test validation

The validation process was guided by Weir's (2005) evidence-based approach. Three types of evidence were gathered, i.e., construct validity, parallel forms reliability, and comparability among the multiple tests. First, we evaluated the construct validity of the reading tests. Construct validity refers to "the extent to which we can interpret a given test score as an indicator of the ability (ies), or construct(s), we want to measure" (Bachman & Palmer, 1996, p. 21). We presented the tests with an evaluation form to two professors in chemistry to evaluate to what extent students' performances on each test could be used as indicators of the six types of reading operations as the construct of "academic reading ability in chemistry". Based on their suggestions, two tests were excluded. Second, parallel forms reliability was evaluated. This reliability estimates the consistency of the difficulty levels of the test forms.

Fifteen undergraduates who had just completed the same course were asked to do the remaining 13 tests, which yielded $15 \times 13 = 195$ test scores. After a reliability analysis based on item-to-total correlation via the SPSS software, 12 tests remained. The average score was 78.00 (Min = 75.1; Max = 80.6; SD = 1.8). The Cronbach's alpha is .83. Finally, to further examine the comparability of the 12 test forms, the Friedman test was used. Results showed no significant difference ($x^2_r = 14.49$, $p = .207 > 0.05$). Based on these three types of evidence, it was suggested that these 12 tests could be used for the current study.

3.3. Setting

The present study was situated in an academic English in Chemistry course at a top-tier comprehensive university in central China for one semester (3 h per week; 16 weeks in total) with 97 undergraduates in their second year. This course, instructed in both English and Chinese, aims to facilitate students' ability to read Chemistry textbooks in English, focusing on discipline knowledge rather than on analyses of reading texts. The instructor was a professor in Chemistry. She obtained her doctoral degree in an English-speaking country and was very proficient in the English language. The textbook used in the course was *Advanced Chemistry through Diagrams* (Lewis, 2001).

3.4. Participants

Thirty-one students from the course participated in the study, 18 females and 13 males, aged from 18 to 20. We obtained informed consent from the participants and compensated them in cash for their time in the study. At the end of the study, four participants were excluded because of incomplete scores, resulting in 27 participants. They were randomly numbered from S1 to S27. Nearly all of them

Table 1
Reading tests information.

Test	Topic	Length	Flesch-Kincaid Grade Level
1	●Mass spectrometer	145	10.8
	●Potential energy and kinetic energy	132	11.9
2	●Sodium chloride crystal	160	13.0
	●Bond dipole and dipole moment	212	10.4
3	●Phase change	295	11.3
	●Alcohol properties factors	175	12.9
4	●Addition polymerization mechanism	201	12.0
	●Natural rubber and synthetic rubber	266	11.8
5	●Diatomic ligands	201	13.0
	●Greenhouse effect	242	12.6
6	●Lithium batteries	325	13.4
	●Fuel cells	239	12.5
7	●Porous carbons (CO_2/N_2 selectivity)	231	12.3
	●Buffer solution	301	9.9
8	●Enzyme under cell regulation	243	11.9
	●Ozone formation and influence	302	13.0
9	●MOFs as proton conductor	231	13.1
	●Electronegativity and reactivity	280	10.5
10	●Reductive amination	227	13.5
	●Entropy and spontaneous gas expansion	279	10.6
11	●Decay of radioisotope	246	11.7
	●Vaporization in open and closed systems	278	11.5
12	●Colligative properties	239	10.2
	●Nuclear fission	319	13.2

had learned English for 10 years. Nine of them were from advanced-level classes and 18 from the intermediate-level classes by a school-based placement test, which consists of listening, reading, and writing tasks. Their average English proficiency was at an intermediate level, roughly at level B2 in terms of the *Common European Framework of Reference for Languages*. They had taken courses in inorganic chemistry, analytical chemistry, and laboratory operation. At the time of this study, they were taking organic chemistry and physical chemistry courses.

3.5. Data collection

English proficiency scores of the participants were based on their examination scores in the College English course in the previous two semesters. The examination was university-based, and was administered to around 7000 college students. It was composed of listening, reading, writing, and integrated items. The test underwent item writing, item screening, item analysis, and test piloting before test administration.

Chemistry knowledge was indicated by the average scores of four courses in chemistry, i.e., two inorganic chemistry courses, one organic course, and one analytical chemistry course. These were the main chemistry courses which they took before the present study. These four courses in discipline knowledge were instructed in Chinese. The maximum possible score for each of them was 100 points, which is a commonly used scoring method in China.

Data collection for the present study started in the first week of the semester and continued for 12 weeks. Each week, the participants were required to do one pen and paper test, which lasted for 30–40 min. Immediately after they completed the reading comprehension test, they were interviewed. The interview was conducted in Chinese. Each interview session lasted for 10–20 min on an individual basis to prevent possible peer influence. After the interview, scores of the previous test were given to the participants, together with answers to the questions on the test. Participants were encouraged to seek help from the researchers if they had English language related questions and discuss with their peers and the researchers about how to improve their reading ability.

3.6. Data analysis

Several analyses were conducted to explore answers to the five research questions.

(1) To investigate the extent to which English ability and Chemistry knowledge, predict the initial academic reading score, a multiple regressions analysis was conducted, with initial score as the dependent variable and English and Chemistry as independents variables.
(2) To measure gains we used the Wilcoxon signed-rank test with the average of the first two tests as beginning score and the average of the last two tests as final score.
(3) To examine to what extent degrees of variability and phase shifts are related to gains in reading ability, we used correlational analyses.

Variability was expressed as one overall number that indicates to what extent the scores fluctuate from one moment to the other, taking the time dimension into account. In line with Pettitt (1980) and Taylor (2000), it was operationalized as the standard deviation of differences between the raw scores and its own average on the basis of the preceding difference (SDd). Different from the standard deviation, the SDd takes time order of the raw scores into account. Thus, differences were calculated by: $d_1 = x_1 - \bar{x}$; $d_i = d_{i-1} + (x_i - \bar{x})$. The SDd yielded one indicator of the variability of a set of scores. To see if any learners showed phase shifts (sudden and radical changes), we used the freely available Change Point analyzer tool (Taylor, 2000).

(4) To visualize how participants group together in the beginning scores and flow to different groups in variability and score gains, a Sankey diagram was created with web-based SankeyMATIC. Sankey diagrams are often used to depict how subsets of one category flow to the subsets of another category with the width of the arrows proportional to the flow rate. To see if we could quantify similar patterns of development in sub-groups of learners, we conducted a cluster analysis. Distance matrices based on individual ranking differences in beginning scores, end scores, and degree of variability were used to retrodict clusters of similar trajectories. These three dimensions were examined because they decided the major displayed features of the trajectory. Other factors, such as English and Chemistry, were not incorporated because they might only influence why a trajectory displayed specific features as it did.
(5) Interviews were used to investigate contributing factors in their development. The interviews were semi-structured and conducted in Chinese. Two prompt questions were a) What were the major difficulties when you did this test? b) What strategies did you use to solve these difficulties? The interviews were recorded with notes and were analyzed through a two-cycle coding system (Miles & Huberman, 1994). First, we established preliminary codes based on thorough reading of the participants' responses. Then, we combined related ones by using pattern codes. Data from interviews in weeks 1–3, 4–6, 7–9, 10–12 were combined to avoid redundancy in reporting. The data in each phase were compared between high gainers (top 30%) and low gainers (bottom 30%) to explore possible contributors to academic reading development.

4. Results

4.1. Predicting initial academic reading ability development

The overall multiple regression with the initial reading score as the dependent variable and two independent variables (i.e., English, Chemistry) was significant ($R^2 = 0.53$, $p < .05$). The standardized coefficient of English was 0.45 ($p < .05$) and Chemistry was 0.17 ($p > .05$). The results indicated that English was a significant predictor of the initial reading score, but Chemistry was not a significant predictor.

4.2. Academic reading gains

The average score of the first two tests was 70.6 points out of 100 and the last two tests 83.4. The average improvement of the 27 participants was 12.8. The Wilcoxon signed-rank test showed that the end score was significantly higher than the beginning score ($Z = 4.21$; $p < .01$; effect size $r = 0.81$).

4.3. Longitudinal development SDd and phase shifts

Each of the 27 participant's developmental trajectory was plotted on a line graph. Fig. 1 shows that all trajectories except two (S4 and S19) were in an ascending trend. The highest point was in the route of S19 with a score of 95 points, and the lowest point was in the trail S22 with a score of 45 points.

For the 27 participants, the correlation between SDd and score gains was significant ($r = 0.68$, $p < .01$) and of a large size according to Plonsky and Oswald (2014). Phase shifts were detected in five participants, namely S5, S6, S18, S22, and S24. As Table 2 indicates, these learners belonged to the top 12 students in terms of gain scores and had relatively high SDd's. The point-biserial correlation analysis revealed that the correlation between the SDd and whether the development of a learner experienced phase shift was significant and of large size ($r = 0.71$, $p < .05$).

The five participants who experienced phase shifts demonstrated different development patterns. S5's developmental trajectory was characterized by alternation of small and big variations (see Fig. 2a). From week 1 to week 4, S5's reading scores were low and the fluctuation was small. S5 experienced the greatest oscillation around week 5. After the regression from 92 to 81, transitory stability was identified around week 10. However, it was soon replaced by greater variation, which persisted into week 12. A phase shift was detected in week 5 (see Fig. 2b).

The developmental trajectory of S6 demonstrated less variation in the initial and final stages, with greater fluctuation in between, revealed by the changing bandwidth (see Fig. 2c). Fluctuation around week 5 was the largest. The change-point analysis detected a sudden change in week 5 (see Fig. 2d). The development of S18 featured small changes in the beginning and sudden increase in the middle (see Fig. 2e). A phase shift was identified in week 7 (see Fig. 2f). S22's trajectory was patterned by a turbulent initial stage and a

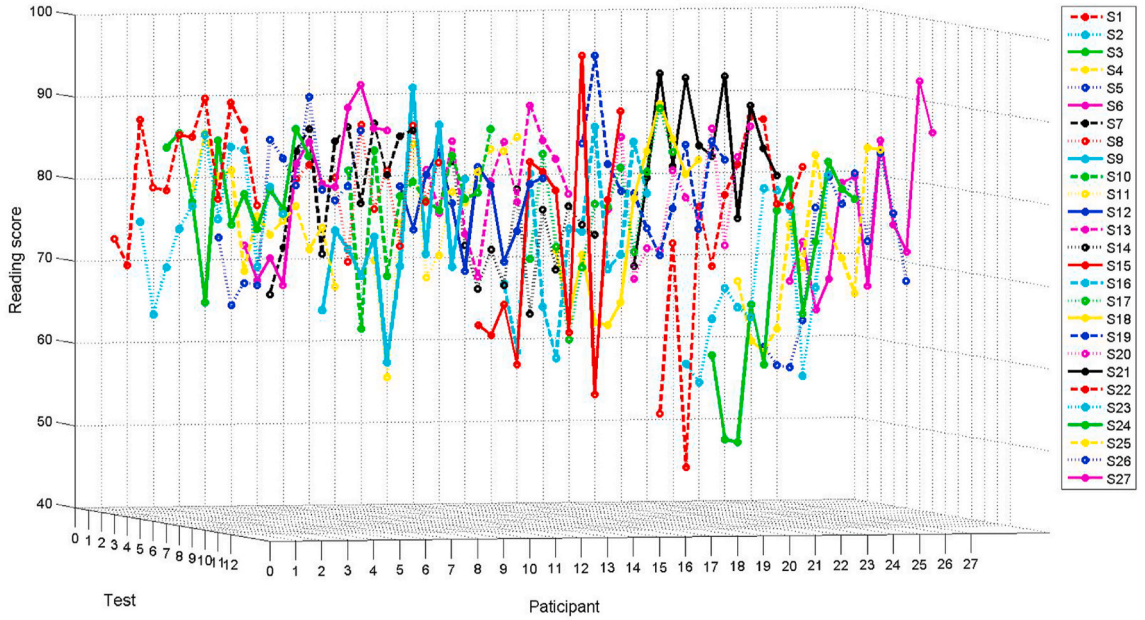


Fig. 1. Developmental trajectories of 27 individuals over 12 weeks. From left to right, they represent developmental trajectories of S1 to S27, respectively.

Table 2
Scores and rankings.

Participant	Score gains	Beginning score (ranking)	Ending score (ranking)	SDd (ranking)	English	Chemistry
S24	27.5	53.0 (27)	80.5 (21)	20.43 (01)	67	68
S11	24	63.0 (21)	87.0 (06)	11.20 (07)	86	73
S15	24	61.5 (23)	85.5 (10)	13.00 (05)	81	76
S25	22.5	63.5 (20)	86.0 (08)	12.78 (06)	76	67
S27	21.5	69.5 (14)	91.0 (01)	11.04 (07)	81	77
S16	21	63.0 (21)	84.0 (15)	11.38 (07)	80	63
S23	20	56.0 (26)	76.0 (25)	20.62 (01)	78	82
S22	20	61.5 (23)	81.5 (20)	12.02 (06)	80	82
S7	19.5	69.0 (16)	88.5 (03)	8.22 (10)	83	79
S6	19	70.0 (13)	89.0 (02)	15.37 (04)	85	78
S18	17	67.0 (19)	84.0 (15)	19.17 (02)	78	86
S5	16.5	69.0 (16)	85.5 (10)	12.96 (06)	88	78
S26	16	58.0 (25)	74.0 (26)	17.95 (03)	74	67
S10	13.5	71.5 (11)	85.0 (12)	7.39 (11)	85	73
S1	13	71.5 (11)	84.5 (13)	9.55 (09)	87	87
S17	12	76.5 (08)	88.5 (03)	11.77 (07)	78	75
S2	11	69.5 (14)	80.5 (21)	10.15 (08)	90	84
S21	10	74.5 (10)	84.5 (13)	7.48 (11)	81	86
S9	8.5	69.0 (16)	77.5 (23)	11.04 (07)	91	78
S20	6.5	80.5 (05)	87.0 (06)	8.23 (10)	79	84
S12	6	76.5 (08)	82.5 (18)	3.65 (12)	90	70
S13	4.5	78.5 (06)	83.0 (17)	8.55 (10)	84	72
S3	2.5	85.0 (02)	87.5 (05)	7.57 (11)	95	83
S8	1.5	81.0 (04)	82.5 (18)	3.89 (12)	82	79
S14	−0.5	77.0 (07)	76.5 (24)	6.03 (12)	82	57
S19	−3.5	89.5 (01)	86.0 (08)	9.36 (09)	79	85
S4	−9	82.5 (03)	73.5 (27)	5.08 (12)	92	82

Note. The students in grey had a phase shift.

more stable later stage (see Fig. 2g). A phase shift was identified around week 4 (see Fig. 2h). S24's trajectory was patterned by continual large variation (see Fig. 2i). One phase shift was detected in week 6 (see Fig. 2j).

4.4. Cluster retrodiction

Fig. 3 shows the Sankey diagram. The first column represents the beginning score, the middle column variability, and the last column the gains. A, B, C, and D were the rankings in each category. The width of each flow was proportionate to the number of students in each category. The diagram shows that students with higher beginning scores (A and B) underwent lower level of variability (C and D). Students experienced higher variability had bigger gains. These findings were in line with the CDST perspective.

In order to test the CDST assumption that no learners are alike, we tested to what extent we could actually group learners into ergodic ensembles and conducted a cluster analysis according to developmental patterns. Table 2 presents the beginning and end scores, gains, SDd, their rankings, as well as their initial English and Chemistry scores. The difference between any two participants' rankings was calculated and used to indicate their distance. Based on the rankings, three distance matrices were created, namely, (a) the distance matrix of beginning scores, (b) the distance matrix of ending scores, and (c) the distance matrix of SDd.

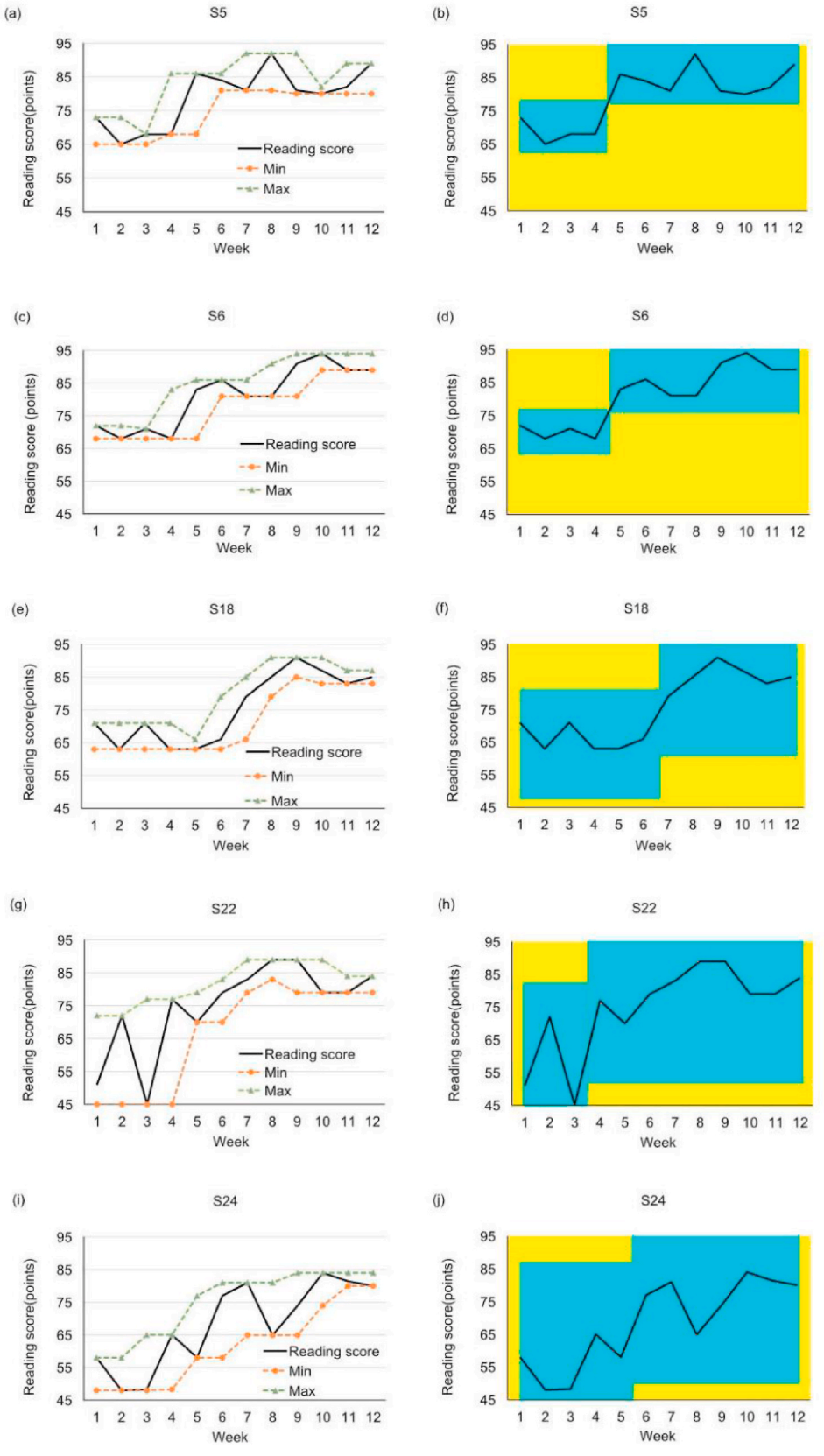
By adding the distances in the three matrices, a sum of distances matrix was constructed (see Appendix). For instance, regarding beginning reading scores, the distance between S1 and S5 was 5 (16th – 11th), and the distance between S1 and S10 was 0 (11th – 11th). Regarding ending scores, the distance between S1 and S5 was 3 (13th – 10th), the distance between S1 and S10 was 1 (13th – 12th). For the SDd, the distance between S1 and S5 was 3 (9th – 6th), the distance between S1 and S10 was 2 (|9th–11th|). Therefore, the sum of distances between S1 and S5 was 11 (5 + 3 + 3), and between S1 and S10, the sum was 3 (0 + 1 + 2). It can be inferred that the trajectory of S1 was more similar to that of S10 than to S5. The smaller the sum of distances was, the more similar the two trajectories were.

Based on information in the Appendix, seven clusters were formed, which encompassed 19 of the total 27 participants (70%). The remaining 8 participants did not fit into a cluster. The seven clusters were listed in the order of score gains. The first cluster with score gains of 23.8 points out of 100 consisted of two participants, S22 and S24. Table 3 lists the score gains of the seven clusters, as well as the participants in each cluster, the beginning and ending score, and variability.

In cluster 1, the trajectories of S22 and S24 were very similar. Their scores were very close most of the time except in weeks 2 and 8 (see Fig. 4a). They both experienced big ups and downs before week 9, and they were stable in the last three weeks. They both improved their reading scores with more than 20 points. They both experienced phase shifts.

The score gains for cluster 2 were 19. The trajectories of S16 and S18 were nearly the same except in week 9 (see Fig. 4b). Trajectories of S16 and S18 were not stable in any period. Their routes shifted much in the first 5 weeks and ascended in the middle weeks. Their English scores were very close, but their Chemistry scores were far apart.

Cluster 3 (S6, S7, S17, and S27) improved 18 points. This cluster varied less (see Fig. 4c). Their English and Chemistry were similar



(caption on next page)

Fig. 2. Trajectories of five cases with phase shifts on the moving min-max and the change-point graphs. Shifts of blue areas with different upper and lower limits suggest a sudden change of the mean of reading scores. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

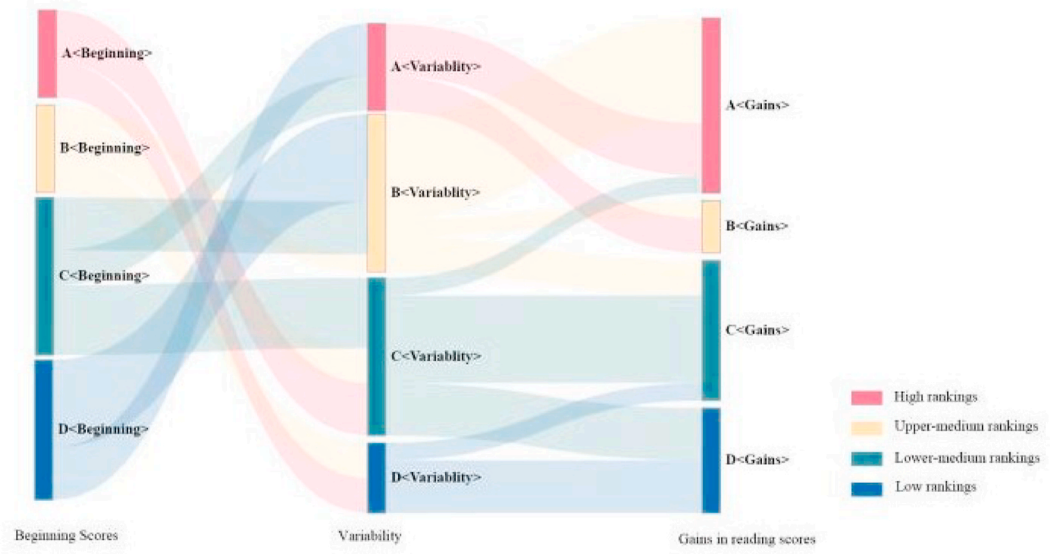


Fig. 3. Flow of students at different rankings of beginning scores to the subsets of variability, and gains in reading.

Table 3
Cluster information.

Cluster	Participant	Gains	Beginning	Ending	SDd	English	Chemistry
1	S22/S24	23.8	57.3	81.0	20.53	73.5	75.0
2	S16/S18	19.0	65.0	84.0	15.28	79.0	75.5
3	S6/S7/S17/S27	18.0	71.3	89.3	11.60	81.8	77.3
4	S23/S26	18.0	57.0	75.0	14.99	76.0	74.5
5	S1/S10/S21	12.2	72.5	84.7	8.14	84.3	82.0
6	S8/S12/S13	4.0	78.7	82.7	5.36	85.3	73.7
7	S3/S19/S20	1.8	85.0	86.8	8.38	84.3	84.0

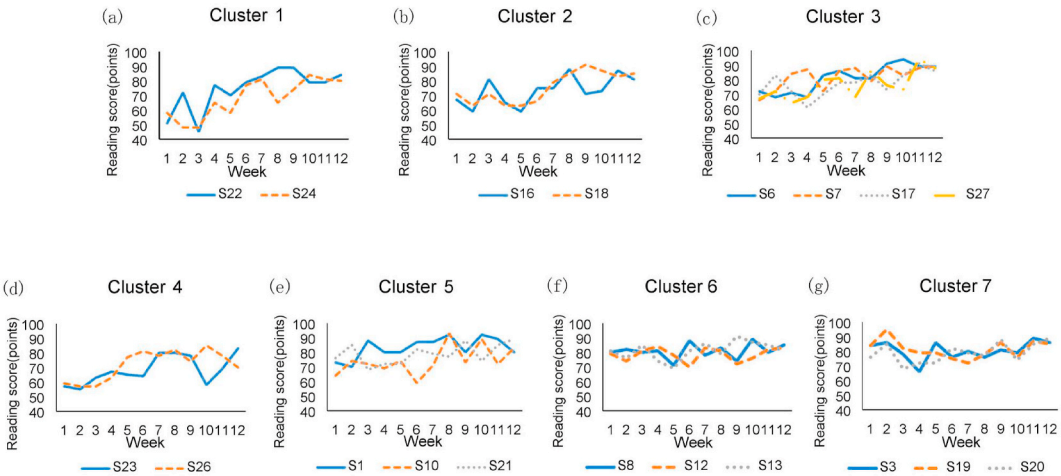


Fig. 4. Trajectories of the seven clusters.

too. The trajectories of S23 and S26 were nearly identical except in weeks 6 and 10 (see Fig. 4d). They were similar in English but different in Chemistry. In cluster 5, the trajectories of S1, S10, and S21 were very close except in week 6 as shown in Fig. 4e. The variability of this cluster was much smaller than that of the first 4 clusters and was similar to those of clusters 6 and 7. They all progressed in a fluctuating manner. All three members in this cluster had good English and Chemistry background.

Clusters 6 (S8, S12, and S13) and cluster 7 (S3, S19, and S20) shared similarities in that they both had small score gains, lower than 5 points, and little variability, but cluster 6's beginning and end scores were lower than cluster 7. The three trajectories in each cluster nearly stuck together (see Fig. 4f and g). Participants in both groups progressed very little although their beginning scores were high. They also had high English and Chemistry scores.

To summarize, the retrodictive clustering was based on similarities in the individual's trajectory in three dimensions. Collectively, the individuals in each cluster were similar in displayed patterns of development, but never exactly the same. The differences between any two clusters in three dimensions were examined. Some clusters were similar in one or two dimensions but no two clusters were close enough in all three dimensions. For instance, clusters 3 and 4 had the same score gains (18 points), but their beginning scores were significantly different. Cluster 3's starting score ($M = 71.3$, $SD = 3.52$) was significantly higher than that of cluster 4 ($M = 57$, $SD = 1.41$) ($t_{[4]} = 5.25$; $p < .05$; Cohen's $d = 5.25$). Therefore, clusters 3 and 4 were considered as two different trajectories.

4.5. Interview results

After every test, the researchers interviewed the students to ask what the major difficulties had been in the test taken that day and what strategies they had used since the previous test to solve the difficulties. Their responses were categorized by using pattern codes. Because the individual data was not structured enough to make exact counts for strategies used, we combined the strategies for two groups, the highest gainers (S24, S11, S15, S25, S27, S16, S23, S22, and S7) and the low gainers (S9, S20, S12, S13, S3, S8, S14, S19, and S4), and show them in four phases of three weeks each (weeks 1–3, 4–6, 7–9, and 10–12).

In the first phase, members from both groups repeatedly and almost unanimously reported that technical vocabulary was the difficulty in reading (see Table 4). Their strategies to deal with the difficulty were also similar, which included memorizing new words and guessing the meaning of unfamiliar words.

However, in the second phase, diverging patterns between high gainers and low gainers began to emerge. High gainers talked about more types of strategies with higher frequency than the lower gainers to memorize technical words, including organizing technical words with tables, organizing technical words with the number of carbons and functional groups, and reading aloud technical words. For example, they arranged words beginning with hex- (meaning six) together, *hexane, hexyl, hexene, hexyne, hexanol,* and *hexanal.*

In the third phase, high gainers mentioned five types of difficulties, while low gainers reported three. Regarding strategies both groups reported that they parsed sentences when they read complex sentences. However, high gainers also talked about strategies of summarizing and analyzing logical relationships between sentences and paragraphs, which were not reported by the low gainers.

Table 4
Interview summary.

Weeks		High gainers ($n = 9$)	Low gainers ($n = 9$)
1–3	Difficulty	technical words (24)[a]; chemistry knowledge (4); complex sentences (2)	technical words (22) complex sentences (9) logic between para (2)
	Strategy	memorize words (17); reread (10); guess word meaning (10); use grammar (2); use chemistry knowledge (3); skip unfamiliar words (4)	guess word meaning (16); memorize words (10); skip (8) scan to find info to questions (5); use grammar (3)
4–6	Difficulty	complex sentences (15); technical words (9); sentential logic (4); summary (2)	technical words (17); complex sentences (15); summary (3)
	Strategy	organize technical words with tables (15); organize technical words with the number of carbons and functional groups (10); ask help from instructors (10); discuss with peers (9); read aloud technical words (9); collect long sentences and restudy them (8); parse sentences (6); use examples to help understanding (3); translate into Chinese (2)	memorize words (10) guess word meaning (8); parse sentences (7); underline keywords (5); organize technical words with tables (3); scan to find answers to questions (5); reread (2)
7–9	Difficulty	sentential logic (9); discursive logic (6); complex sentences (6); technical words (4); summary (4)	technical words (15); complex sentences (10); summary (7)
	Strategy	parse sentences (15); summarize (6); analyze the logic between paragraphs (5); analyze the logic between sentences (5); collect complex sentences and restudy them (4); adjust reading speed (2)	memorize words (10) parse sentences (10); guess word meaning (3); translate into Chinese (3); skip unfamiliar words (2)
10–12	Difficulty	summary (10); discursive logic (9); sentential logic (8); chemistry knowledge (2); answer questions correctly (2)	discursive logic (9) sentential logic (9); technical words (8); summary (5)
	Strategy	summarize (10); identify key info (7); increase reading speed by skipping examples (4); parse sentences (4); check understanding (3); draw graphs to study relation between paragraphs (2); use chemistry formula to understand the main idea (1)	parse sentences (10); summarize (5); analyze logic between paragraphs (5); analyze logic between sentences (5); locate key information (3)

[a] The number refers to frequency of one type of reading difficulty or strategy mentioned by all members in a group during a period of three weeks.

In the last phase, high gainers mentioned difficulty in summarizing, understanding sentential and discursive logic, whereas low gainers consistently reported on their difficulties in technical vocabulary along with sentential and discursive problems. High gainers reported seven types of strategies while low gainers mentioned five. Some unique strategies for high gainers included drawing graphs to understand logical relations between paragraphs.

The interview data suggest that high gainers' comments on difficulty and strategy varied from those of low gainers. Although they met similar difficulties, high gainers reported more specific, detailed, and creative strategies than the low gainers. For example, to cope with technical vocabulary, S7, S15, and S24, from the high gainer group, organized vocabulary according to the number of carbons and functional groups and read them out aloud. In contrast, S14 and S19 from the low gainer group also talked about the need to memorize more technical words, but they did not mention any attempts to help memorization except for rote memory. It is very likely they did not invest time or effort in taking actions to tackle the problem.

To solve the problem of complex sentences, S24, who had the highest score gain, reported that she gathered all long and complex sentences in the passages of the previous tests and restudied them. She also checked with the researchers of the present study to ensure she understood the sentences correctly. To cope with difficulty in summary, S15 reported that he summarized a passage with a chemistry formula or a graph. However, such detailed and creative strategies were seldom reported among low gainers.

To summarize, high gainers were more engaged in the study and specific, detailed, creative, and more diverse and pro-active in their use of strategies, whereas low gainers kept on talking about the technical problems and complex sentences, probably because they were not too motivated and did not invest enough time in taking actions to solve the problems.

5. Discussion and implications

In a mixed method approach, this study traced the academic reading development of 27 Chinese Chemistry majors longitudinally from a CDST perspective. Twelve validated parallel versions of an academic reading test were used during one semester and after each test the participants were interviewed on what they found the most difficult in the reading test and what strategies they had developed since the last test and what they would focus on for the next test. The findings and implication for each research question will be dealt with separately.

5.1. To what extent can initial academic reading score gains be predicted by English proficiency or chemistry knowledge?

In this particular group, English proficiency was a significant predictor of the initial reading score, but Chemistry was not. The fact that English proficiency played a strong role is not surprising as findings are consistent especially with Usó-Juan (2006) and Davis, Huang, & Yi, 2017 where language proficiency predicted EAP reading more than domain knowledge. It was surprising though that Chemistry did not play a significant role at all as in Usó-Juan (2006) and Davis, Huang, & Yi, 2017. This could be explained by the fact that all participants already had domain knowledge of Chemistry as they were Chemistry majors and even though some had a higher average grade for their Chemistry courses in the semester before (our predictor), all students probably had enough basic domain knowledge to understand the Chemistry aspects in the text.

5.2. Does the group gain in academic reading scores after one semester?

As a group, the Chinese Chemistry major undergraduates' academic reading ability developed significantly. It was found that nearly all participants' academic reading ability developed in an ascending pattern with ups and downs during the study period, with the greatest up and downs in the middle of semester and a stabilization towards the end. Of course, it is no surprise that the scores improved on the whole, as the tests themselves were excellent learning opportunities and are in line with Pritchard and Nasr (2004), who recommend using authentic materials that are used to comprehend concepts is effective in improving academic reading ability. The texts we used were indeed authentic tests and the questions focused on understanding the content of the texts, which included domain knowledge. Moreover, the retrospective interviews may have helped in making students aware of where they were weak, what their problems were and as most students obviously wanted to improve, they sought out ways to improve.

5.3. To what extent does variability in academic reading scores (in terms of SDd and phase shifts) play a role in gains?

Within CDST it is assumed that individuals will create their own learning trajectories and will show non-linear behavior resulting in *variability*, which is assumed to be needed to progress, and thus *variation*, in that no two learners are exactly alike.

CDST inspired studies so far have focused especially on L2 writing development and authors acknowledge that variability could also be partly attributed to writing task or topic effects (cf. Lowie & Verspoor, 2019). The contribution of the present study is that the reading tests had been calibrated and were assumed to be highly similar. Therefore, the variability was not because of the task or topic, but clearly inherent in the developing learners. This is a clear confirmation of earlier CDST findings.

And also in our study, we found that variability over time (operationalized as the SDd) was highly correlated with reading ability gains. This result was in line with what has been observed in Lowie and Verspoor (2019). The correlation in our study ($r = 0.68$) was higher than that in their study ($r = 0.53$). We also investigated whether phase shifts occurred in the course of any individual's development because the CDST perspective pays particular attention to the sudden and radical changes (Baba & Nitta, 2014; Larsen-Freeman & Cameron, 2008). It was found that 5 among 27 participants (19%) experienced phase shifts. The SDd was also significantly correlated with phase shifts. When we look at Table 2, we can see that the lowest 30% of learners had low levels of

variability and no phase shifts. These findings suggest that even though there may not be a one-on-one relation, variability and phases-shifts do not occur in learners who do not progress or even regress. In other words, learners will find their own unique trajectories in development, and a degree of variability is needed to improve. As Huang, Steinkrauss, and Verspoor (2021) point out, variability is only a symptom of behavior, and the interview data imply that the high gainers seem to be more motivated to improve, more active to seek help from the researchers of the study, and more creative in thinking of strategies to deal with difficulties in reading. These results suggest that phase shifts may be important, but by looking only at phase shifts, we may miss other interesting developmental patterns, such as overall variability over time, which is also a sign of progress.

5.4. Are there groups of individuals who share similar developmental trajectories?

To test the CDST assumption that no learners are alike, we first visualized group patterns with a Sankey diagram. The diagram showed that in general students with the lowest initial scores had relatively more variability and gained the most. On the other hand, students with the highest beginning scores had relatively less variability and gained the least. However, the bandwidths in the flow chart are large, so there is variation among learners. To see if we could detect similar "ergodic ensembles", we also clustered trajectories with similar patterns based on their displayed performances. Seven clusters were formed with 2–4 participants in each and 19 participants (70%) in total. Although the importance of cluster analyses from a CDST perspective has been suggested by some researchers (e.g., Dörnyei, 2014; Lowie & Verspoor, 2019; Zheng, 2018, pp. 218–229), few previous studies have been able to present effective quantitative approaches.

The difficulty lies in that members in each cluster need to be "the most similar in as many respects as possible" (Lowie & Verspoor, 2019, p. 195) while traditional approaches can hardly deal with such multiple dimensions. The strength of using distance matrices is that distance between any two participants in one dimension can be calculated and overall distance can be obtained by adding them up. Regarding the number of dimensions that should be taken into account, the researcher needs to estimate a tradeoff between the degrees of member homogeneity desired and the usefulness of clustering. In our case, only 19 of the 27 participants could be clustered, but even then, members of the clusters differed in one or more of the dimensions. These findings suggest that there is definitely variation among learners, as each has his or her own individual developmental trajectory, illustrating that even small groups of humans are not strong ergodic ensembles. These findings confirm CDST assumptions of variation among learners.

5.5. What factors do students comment on in terms of difficulties and strategies, and can we relate them to academic reading score gains?

Factors that influenced score gains were further explored by using semi-structured interviews. The results revealed that the possible factors included memorization of technical vocabulary, increased sentential and discursive ability, and summarizing ability. What is interesting to note is that almost all students started off with vocabulary learning strategies in Phase 1, which is understandable because a coverage of 95%–98% of the words (Schmitt, Cobb, Horst, & Schmitt, 2017) is needed to be able to understand a text. Confronted with similar difficulties, the low gainers reported limited and general strategies, mostly rote memory. In other words, they were neither really motivated nor creative in finding strategies to improve.

In contrast, the high gainers were more mindful of their difficulties and more active in finding specific and creative strategies to improve their vocabulary. Moreover, they started to look beyond vocabulary and were aware that complex sentences were a problem and found more ways to understand them better than the low gainers. In phase 3, the low gainers also looked at complex sentences and used a few strategies to improve understanding them, but the high gainers used more detailed and concrete comprehension strategies. In the final phase, the high gainers had adopted more global text strategies and knew better how to adapt them according to context. The low gainers made fewer mention of global reading strategies. Thus, high gainers used more specific, creative, and flexible strategies than the low learners, and they progressed from vocabulary, sentence level, to whole text strategies over time. This is very much in line with Siegler (2006), who suggested that new strategies are added as time progresses, with greater reliance on relatively more advanced strategies, and improved choices among strategies. It is also in line with McDonough (1999) who found that successful learners generally use a greater number and a wider variety of learning strategies. The findings are very much in line with Cai and Kunnan (2019; 2020) who found that some factors such as reading strategies fluctuate with changes in language proficiency.

5.6. Implications

The implications of these findings for researchers are that a combination of traditional cross-sectional, CDST inspired longitudinal, and qualitative data gives a nuanced view of development. Traditional statistical analyses showed that the group as a whole progressed significantly in reading over time and that English proficiency early on was a predictor for the first score. The CDST inspired longitudinal analyses showed that even though the group as a whole progressed, there was much variation among the learners and each learner had his or her own individual trajectory with different degrees of variability. Moreover, we found that generally the higher gainers tended to be more variable and a few of them had clear phase shifts. The qualitative data showed that the number and types of strategies changed over time, with the high-gainers using more sophisticated strategies to improve at the end. Thus, when measuring any construct, be it reading, writing or strategy use, it is useful to have a combination of methods and to have more than one test moment, especially before a learner's L2 system has stabilized.

Several pedagogical implications may be derived from our findings. First, L2 learners usually undergo diverse developmental patterns. They differ almost in each L2 task performance and degrees of change during the whole learning course, especially if they are actually making progress. Therefore, teachers need to be aware of the idiosyncratic features of students and realize that ups and downs

are normal in development. Moreover, students who seem to go up and down the most may actually be learning the most. The more effective learners will probably be autonomous enough to monitor their difficulties in reading and try to overcome them by trying various strategies. The less effective learners do not seem to be conscious enough of their reading difficulties, or they are not determined to cope with them, so the problems remain unsolved, and little progress can be achieved. In such cases, teachers may ask them about their motivation to improve and suggest the use of different strategies. Second, the students' responses in the interviews clearly showed a progression in difficulties and strategies. For this group, the technical vocabulary was the most difficult and not until they had improved enough, did they start looking at ways to improve on reading complex sentences and discursive patterns and logic. In other words, the types of strategies used changed over time, but vocabulary knowledge is the starting point and in teaching academic reading teachers may need to focus on vocabulary first. Finally, even though our experiment was not set up to be an educational intervention, it was surprising to see how much the students improved over time and worked hard to improve, which reminds us of the idea of formative assessment in that it helped students identify their strengths and weaknesses and provided target areas that need work.

The limitations of the study are that it represents the findings of a particular group in a particular context and may not be generalizable to all other contexts. However, from a CDST perspective that is to be expected. There is not only variation among members of a group, but also variation among groups, especially in different contexts. Therefore, it would be useful to replicate this study in other contexts to see if similar developmental patterns occur. Another limitation is that the study was not able to untangle the interaction between domain knowledge and English proficiency over time as they were conflated in one reading score. In future CDST-based research, such an investigation might be worthwhile given that from a CDST perspective interaction among system components and interaction between a system and its environment give rise to the development of a system.

6. Conclusion

Employing a CDST perspective, the present study has investigated the developmental trajectory of Chinese Chemistry major undergraduates' academic reading ability by means of repeated measure over time. After a semester of an academic reading course in which they took a test each week, the group as whole progressed significantly in academic reading. The initial reading scores were very much predicted by their English proficiency. However, none of the learners developed linearly, all showed ups and downs, but at different moments. The cluster analysis showed some sub-patterns but no two learners were alike in all respects. There was a strong correlation between gains and degree of variability and among the 30% low gainers there was little variability and no phase shifts.

Variability is just a symptom of behavior and not a cause or character trait, and we may find some clues in the interview data for these symptoms. The interview data showed that the high gainers, most of whom showed larger degrees of variability, reported specific and creative strategies to improve themselves. They realized where their weaknesses were and explored more facilitative ways than the low gainers to improve their performance. Finally, the difficulties they faced and strategies they thought of changed over time: from vocabulary to sentence level to discursive features of the text. In other words, as Lowie and Verspoor (2019) point out "more variability may be a characteristic of a creative learning process, in which new things are tried out that may go wrong but lead to an exciting process" (p.19).

The present study may contribute to the literature in three aspects. First, it is one of the first CDST-based empirical studies that have focused on L2 academic reading ability with similar tasks. Most CDST-based studies have concentrated on L2 productive abilities in which task effects may have contributed to variability. Second, we have estimated that about 19% participants experienced phase shifts and a higher degree of individual developmental variability over time is significantly correlated with higher academic reading ability gains. These findings are rather new to the SLA field and need further investigation. Is variability a sign of progress, and if so at what stage of the learning process does it occur? Finally, we have employed distance matrices to quantitatively capture similarities in multiple dimensions of individuals' development trajectories. This creative approach might inspire researchers to seek further measures to investigate the challenging issue of clustering individuals of fluctuating variability within themselves and heterogeneous variation between them.

Author statement

Min Gui: Conceptualization, Investigation, Methodology, Writing – original draft, Funding acquisition, Xiaokan Chen: Data curation, Software, Visualization, Validation, Marjolijn Verspoor: Supervision, Analysis, Writing - Reviewing and Editing.

Funding

This work was supported by China Ministry of Education Humanities and Social Sciences Research Fund [18YJA740014].

Role of the funding source

The funding source had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Declaration of competing interest

None.

Acknowledgments

We would like to gratefully acknowledge three professors in Chemistry, Dr. Sixue Cheng, Dr. Tao Cai, and Dr. Zhenlin Zhong for valuable help in the test design and validation, and the anonymous reviewers and the editors for insightful comments and suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.system.2021.102546.

References

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*, 236–260. https://doi.org/10.1093/applin/amt046

Amer, A. A. (1994). The effect of knowledge-map and underlining training on the reading comprehension of scientific texts. *English for Specific Purposes, 13*, 35–45. https://doi.org/10.1016/0889-4906(94)90023-X

Anderson, N. J. (2015). Academic reading expectations and challenges. In N. Evans, N. J. Anderson, & W. Eggington (Eds.), *ESL readers and writers in higher education: Understanding challenges, providing support*. New York: Routledge.

Baba, K., & Nitta, R. (2014). Phase transitions in development of writing fluency from a complex dynamic systems perspective. *Language Learning, 64*, 1–35.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Cai, Y., & Kunnan, A. J. (2019). Detecting the language thresholds of the effect of background knowledge on a language for specific purposes reading performance: A case of the island ridge curve. *Journal of English for Academic Purposes, 42*, 1–13. https://doi.org/10.1016/j.jeap.2019.100795

Cai, Y., & Kunnan, A. J. (2020). Mapping the fluctuating effect of strategy use ability on English reading performance for nursing students: A multi-layered moderation analysis approach. *Language Testing, 37*, 280–304. https://doi.org/10.1177/0265532219893384

Carrell, P. L., & Carson, J. G. (1997). Extensive and intensive reading in an EAP setting. *English for Specific Purposes, 16*, 47–60. https://doi.org/10.1016/S0889-4906(96)00031-2

Carr, T. H., & Levy, B. (1990). *Reading and its development: Component skills approaches*. San Diego: Academic Press.

Chen, Q., & Donin, J. (1997). Discourse processing of first and second language biology texts: Effects of language proficiency and domain-specific knowledge. *The Modern Language Journal, 81*, 209–227. https://doi.org/10.1111/j.1540-4781.1997.tb01176.x

Cheng, A. (2008). Individualized engagement with genre in academic literacy tasks. *English for Specific Purposes, 27*(4), 387–411. https://doi.org/10.1016/j.esp.2008.05.001

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing, 18*, 133–147. https://doi.org/10.1177/026553220101800202

Davis, D., Huang, B., & Yi, T. (2017). Making sense of science texts: A mixed-methods examination of predictors and processes of multiple-text comprehension. *Reading Research Quarterly, 52*, 227–252. https://doi.org/10.1002/rrq.162

Dörnyei, Z. (2014). Researching complex dynamic systems: 'Retrodictive qualitative modelling' in the language classroom. *Language Teaching, 47*(1), 80–91. https://doi.org/10.1017/S0261444811000516

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732911

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing, 18*, 171–185. https://doi.org/10.1177/026553220101800204

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes, 6*(1), 3–17. https://doi.org/10.1016/j.jeap.2006.11.005

Fraser, C. (2007). Reading rate in L1 Mandarin Chinese and L2 English across five reading tasks. *The Modern Language Journal, 91*, 372–394. https://doi.org/10.1111/j.1540-4781.2007.00587.x

Habók, A., & Magyar, A. (2018). Validation of a self-regulated foreign language learning strategy questionnaire through multidimensional modelling. *Frontiers in Psychology, 9*, 1–11. https://doi.org/10.3389/fpsyg.2018.01388

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing, 32*, 317–336. https://doi.org/10.1177/0265532214564505

Hartshorn, K. J., Evans, N. W., Egbert, J., & Johnson, A. (2017). Discipline-specific reading expectation and challenges for ESL learners in US universities. *Reading in a Foreign Language, 29*(1), 36–60.

Huang, T., Steinkrauss, R., & Verspoor, M. (2021). Variability as predictor in L2 writing proficiency gains. *Journal of Second Language Writing*, [100787]. https://doi.org/10.1016/j.jslw.2020.100787

Jackson, J. (2005). An inter-university, cross-disciplinary analysis of business education: Perceptions of business faculty in Hong Kong. *English for Specific Purposes, 24*(3), 293–306. https://doi.org/10.1016/j.esp.2004.02.004

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL (unpublished doctoral dissertation)*. Urbana and Champaign, IL: University of Illinois at Urbana-Champaign.

Jordan, B. (1997). *English for academic purposes*. Cambridge: Cambridge University Press.

Kaewpet, C. (2009). Communication needs of Thai civil engineering students. *English for Specific Purposes, 28*(4), 266–278. https://doi.org/10.1016/j.esp.2009.05.002

Kasper, L. F. (1995). Theory and practice in content-based ESL reading instruction. *English for Specific Purposes, 14*(3), 223–230. https://doi.org/10.1016/0889-4906(95)00012-3

Kuzborska, I. (2011). Teachers' decision-making processes when designing EAP reading materials in a Lithuanian university setting. *Journal of English for Academic Purposes, 10*(4), 223–237. https://doi.org/10.1016/j.jeap.2011.07.003

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*, 590–619. https://doi.org/10.1093/applin/aml029

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. London: Oxford University Press.

Lenzing, A. (2015). Exploring regularities and dynamic systems in L2 development. *Language Learning, 65*, 89–122. https://doi.org/10.1111/lang.12092

Lewis, M. (2001). *Advanced chemistry through diagrams*. Oxford: Oxford University Press.

Lister, T., & Renshaw, J. (2000). *Understanding Chemistry for advanced level*. Cheltenham: Stanley Thornes.

Lowie, W. M., & Verspoor, M. H. (2019). Individual differences and the ergodicity problem. *Language Learning, 69*, 184–206. https://doi.org/10.1111/lang.12324. S1.

Martınez, A. C. L. (2002). Empirical examination of EFL readers' use of rhetorical information. *English for Specific Purposes, 21*(1), 81–98. https://doi.org/10.1016/S0889-4906(00)00029-6
McDonough, S. H. (1999). Learner strategies. *Language Teaching, 32*, 1–18.
McMurry, J., Ballantine, D. S., Hoeger, C. A., Peterson, V. E., & Castellion, M. (2014). *Fundamentals of general, organic, and biological Chemistry*. Boston: Pearson.
Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis*. Thousand Oaks, CA: Sage.
Nergis, A. (2013). Exploring the factors that affect reading comprehension of EAP learners. *Journal of English for Academic Purposes, 12*, 1–9. https://doi.org/10.1016/j.jeap.2012.09.001
Ortega, L. (2011). SLA after the social turn: Where cognitivism and its alternatives stand. In D. Atkinson (Ed.), *Alternative approaches to second language acquisition* (pp. 179–192). New York: Routledge.
Oxford, R. (2017). *Teaching and researching language learning strategies: Self-regulation in context*. New York: Routledge.
Pettitt, A. N. (1980). A simple cumulative sum type statistic for the change-point problem with zero-one observations. *Biometrika, 67*, 79–84.
Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*, 878–912. https://doi.org/10.1111/lang.12079
Polat, B., & Kim, Y. (2014). Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics, 35*, 184–207. https://doi.org/10.1093/applin/amt013
Pritchard, R. M. O., & Nasr, A. (2004). Improving reading performance among Egyptian engineering students: Principles and practice. *English for Specific Purposes, 23*, 425–445. https://doi.org/10.1016/j.esp.2004.01.002
Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), nation (2006) and Cobb (2007). *Language Teaching, 50*(2), 212–226. https://doi.org/10.1017/S0261444815000075
Siegler, R. S. (2006). Microgenetic analyses of learning. In D. Kuhn, & R. Siegler (Eds.), *Handbook of child psychology, volume 2: Cognition, perception, and language* (6th ed., pp. 464–510) (Hoboken: Wiley).
Silberberg, M. (2013). *Principles of general chemistry*. New York: McGraw-Hill.
Snyder, C. H. (2003). *The extraordinary Chemistry of ordinary things. Hoboken*. New Jersey: Wiley.
Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics, 31*, 532–553. https://doi.org/10.1093/applin/amq001
Taylor, W. A. (2000). Change-point analysis: A powerful new tool for detecting changes. http://www.variation.com/cpa/tech/changepoint.html. (Accessed 29 November 2018).
Usó-Juan, E. (2006). The compensatory nature of discipline-related knowledge and English-language proficiency in reading English for academic purposes. *The Modern Language Journal, 90*, 210–227. https://doi.org/10.1111/j.1540-4781.2006.00393.x
Ward, J. (2001). Est: Evading scientific text. *English for Specific Purposes, 20*(2), 141–152. https://doi.org/10.1016/S0889-4906(99)00036-8
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave MacMillan. https://doi.org/10.1075/lllt.29
Weir, C. J., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2012). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. In L. Taylor, & C. J. Weir (Eds.), *IELTS collected papers 2: Research in reading and listening assessment*. Cambridge: Cambridge University Press.
Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English. *System, 56*, 40–53. https://doi.org/10.1016/j.system.2015.11.007
Zheng, Y. (2018). *'Gao shui ping xue xi zhe yu yan fu za du de duo wei fa zhan yan jiu [The multidimensional development of advanced learners' linguistic complexity],' Wai Yu Jiao Xue Yu Yan Jiu (Foreign Language Teaching and Research)* (vol. 50).