2018

Improving the validity of Script Concordance Testing by optimising and balancing items

Michael Wan

The University of Notre Dame Australia, michael.wan@nd.edu.au

Elina Tor

The University of Notre Dame Australia, elina.tor@nd.edu.au

Judith Nicky Hudson

This article was originally published as:
Wan, M., Tor, E., & Hudson, J. N. (2018). Improving the validity of Script Concordance Testing by optimising and balancing items. *Medical Education, 52*, 336-346.

Original article available here:
http://dx.doi.org/10.1111/medu.13495

Improving the validity of Script Concordance Testing

Improving the validity of Script Concordance Testing by optimising and balancing items.

Siu Hong WAN[1], Elina Tor[1], Judith N Hudson[2]

[1]School of Medicine, University of Notre Dame. Australia

[2]Adelaide Medical School, University of Adelaide. Australia

Email: michael.wan@nd.edu.au

Abstract

Introduction

Script Concordance Testing (SCT) is a modality for assessing clinical reasoning. Concerns had been raised about the plausible validity threat to SCT scores if students deliberately avoided the extreme answer options to obtain higher scores. The aims of the study were firstly to investigate whether student avoidance of the extreme answer options could result in higher scores, and secondly to determine whether a 'balanced approach' by careful construction of SCT items (to include extreme as well as median options as model responses) would improve the validity of the SCT.

Improving the validity of Script Concordance Testing

Methods

Using the paired sample *t*-test, the actual average student scores for ten SCT papers from 2012-2016 were compared with simulated scores. The latter were generated by recoding all ‘-2’ responses to ‘-1’ and ‘+2’ responses to ‘+1’ for the whole and bottom 10% of the cohort (simulation 1), and scoring as if all students had chosen ‘0’ for their responses (simulation 2). The actual average and simulated average scores in 2012 (before the ‘balanced approach’) were compared to those from 2013-16, when papers had a good balance of modal responses from the expert reference panel.

Results

In 2012, a score increase was seen in simulation 1 in the third year cohort; from 50.2% to 55.6% (*t* (10)=4.818; p=0.001). Since 2013, with the ‘balanced approach’, the actual SCT scores (57.4%) were significantly higher than scores in both simulation 1 and 2 (46.7% and 23.9% respectively).

Conclusions

Improving the validity of Script Concordance Testing

When constructing SCT examinations, apart from the rigorous pre-examination optimisation, it is desirable to achieve a balance between items which attracts extreme, as well as median response options. This could mitigate the validity threat to SCT scores, especially for the low performing students who have previously been shown to only select median responses and avoid the extreme responses.

Keywords: medical education, script concordance testing, assessment.

Improving the validity of Script Concordance Testing

Introduction

Script Concordance Testing (SCT) is a modality for assessing clinical reasoning and data interpretation skills in the context of uncertainty. The SCT, introduced in 2000 by Charlin, aimed to assess the higher order clinical reasoning skills of medical students (1).

In a classical SCT question, a clinical scenario is presented in the question stem and the students are then asked to assess whether an additional piece of information increases or decreases the probability of the proposed diagnosis, investigation or management on a 5 point Likert scale. In a SCT question looking at the probability of a diagnosis, for example, if the additional information makes the probability of the diagnosis much more likely, the student will choose '+2'; more likely a '+1'; neither less nor more likely a '0'; less likely a '-1'; and much less likely a '-2' respectively. A sample SCT question is shown in Table 1. Each question under the same clinical scenario is intended to be independent of the other questions, that is, each additional piece of information is not influencing the probability of the diagnosis in the other questions. For Q3 in the sample question, in thinking of the diagnosis of carcinoma of the colon, the student will not consider the additional information of a normal blood glucose or TSH level (2). Although it can be hard for examinees to simply "disregard" previous hypotheticals and data in each question, they are reminded of the need to do this during the pre-examination briefing.

Table 1.

Sample SCT questions:

<table>
<tr><td colspan="5"><u><b>Clinical Scenario</b></u>
<b>A 45-year-old woman presents to the GP clinic with weight loss of 5kg in 2 months. She has no significant past medical history.</b></td></tr>
<tr><td></td><td>If you were thinking of...</td><td>and then you find that…</td><td>this hypothesis becomes …</td><td rowspan="4"><b>-2 :</b> much less likely
<b>-1 :</b> less likely
<b>0 :</b> neither more nor less likely
<b>+1 :</b> more likely
<b>+2 :</b> much more likely</td></tr>
<tr><td><b>Q1.</b></td><td>Diabetes mellitus</td><td>Normal fasting blood sugar</td><td><b>A B C D E</b>
<b>-2 -1 0 +1 +2</b></td></tr>
<tr><td><b>Q2.</b></td><td>Graves' disease</td><td>Normal TSH level</td><td><b>A B C D E</b>
<b>-2 -1 0 +1 +2</b></td></tr>
<tr><td><b>Q3.</b></td><td>Carcinoma of the colon</td><td>A normal digital rectal examination</td><td><b>A B C D E</b>
<b>-2 -1 0 +1 +2</b></td></tr>
</table>

Improving the validity of Script Concordance Testing

To score the SCT items, the student's selection is compared to the decision of an expert clinician panel. A full mark will be given if the student's response is in concordance with the majority of the panel (that is the panelist's modal response). A partial score will be awarded if the response is in concordance with the minority and a zero score for a response that no panelist had selected. An example of the scoring system is shown in Table 2 (3). A minimum of 10 and preferably 15 clinicians would make the scoring process more reliable (4).

Table 2: Formula to calculate the weighted scores in the SCT.

Response Options	**-2**	**-1**	**0**	**+1**	**+2**
Number of clinicians choosing the answer (out of 10)	7	2	1	0	0
Formula	7/7	2/7	1/7	0/7	0/7
Student's score	1	0.29	0.14	0	0

The SCT has been used in undergraduate medical schools examinations as well as in post graduate fellowship training. Successful implementation of SCT has been documented in Medicine, Surgery, Paediatrics, Emergency Medicine, Anaesthesia, Psychiatry and Ethics (5-10). SCT has also been used to assess clinical reasoning in other health care professions e.g. Optometry and Physiotherapy (11, 12). There is evidence of SCT validity and reliability in the literature (9, 13-15) but these remain issues of ongoing debate (16, 17). The reliability of SCT scores has been reported to be around a Cronbach’s alpha value of 0.7-0.85 (9, 14). The construct validity of SCT has been shown by various studies demonstrating progression of SCT scores from undergraduate medical students to post graduate fellows in training (8, 15, 18-20).

However, a recent study has suggested that the aggregate partial credit scoring method used in SCT could be subjected to validity threats (17). Lineberry et al showed students who avoided selecting the extreme response options (i.e. ‘-2’ or ‘+2’), as a ‘*strategic*’ answering approach outperformed other examinees who used the Likert scoring scale as it was intended (consideration of all response options). In their study involving a selected SCT test of 40 items, these authors found that by simulating the avoidance of extreme response options and recoding all responses of ‘-2’ and ‘+2’ to ‘-1’ and ‘+1’ respectively; a phenomenon they called “score inflation” was observed i.e. the hypothetical examinees’ mean score increased from 49.5% to 69.2%. In the same test, a hypothetical examinee who only choose to answer ‘0’ to all items would score 57.6% which would be 8% more than the cohort mean score not using this strategy (17). This significant increase in the examinees’ scores is similar to the response style coaching strategies described in situational judgment tests, which also use a partial aggregate scoring approach (21, 22).

In another study using 2 sets of 96-item SCTs in pulmonary and critical care for post-graduate trainees, simply avoiding extreme answers boosted the Z-scores of the lowest 10 scorers on both SCT sets by ≥ 1 SD (23). The author concluded that increasing the proportion of SCT items with extreme response options (i.e. '+2' and '-2') would attenuate the potential benefit in scores from adopting an "avoidance of extreme responses approach".

Earlier research has revealed that students whose SCT scores were in the lowest quartile were more likely to avoid the extreme answer options in answering SCT questions (24). Given this finding and prior research demonstrating that students who only selected median responses could potentially achieve SCT examination scores that reflected their test-taking, rather than clinical reasoning abilities (17), further research was warranted to test this hypothesis in another setting.

Aims

The study had the following three aims:

1. To investigate whether avoiding the extreme options in SCT tests would result in an increase in the average SCT scores for the whole cohort, and/or for the bottom

10% of cohorts.

2. To investigate through a simulated scoring activity, the outcome of examinees who select only the ‘neutral’ options (‘0’).
3. To determine whether use of pre-examination test optimisation by selecting a logical ‘ideal’ response pattern, and careful construction of SCT items (to include items with both extreme as well as median responses as the modal responses from the expert reference panel) would reduce the likelihood of students benefiting from a potential ‘strategic answering approach’, and improve the validity of the SCT scores.

Methods

Preparation of SCT items and expert panel responses

The School has been implementing SCT in the summative examinations for Year 3 and Year 4 clinical year students in the four-year graduate-entry medical program since 2012. Each SCT examination contains 40 SCT items incorporated as the second part of a multiple choice written paper. The SCT examination in each year covers different disciplines aligning with the specialty teachings relevant to the year. For example, Paediatrics, Psychiatry, Medicine and Obstetrics & Gynaecology in Year 3; Anaesthesia, Emergency Medicine and Surgery in Year 4. Therefore, the SCT questions are different between Year 3 and Year 4. After the Year 3 summative examination, no specific feedback is given to the students as the database of summative SCT questions is limited. However practice SCT questions with the respective expert clinician panel scores are provided in the mid-year as a formative examination to each year.

The expert clinician reference panel consists of practising specialists and general practitioners who are currently involved in the teaching and supervision of the students. The number of clinicians in the panel ranges from 15 to 20 depending on the year of the examination. As previously reported, to set the past/fail score of each SCT examination we used the expert reference panel's mean minus 4 standard deviations (SD) as the cut score (7).

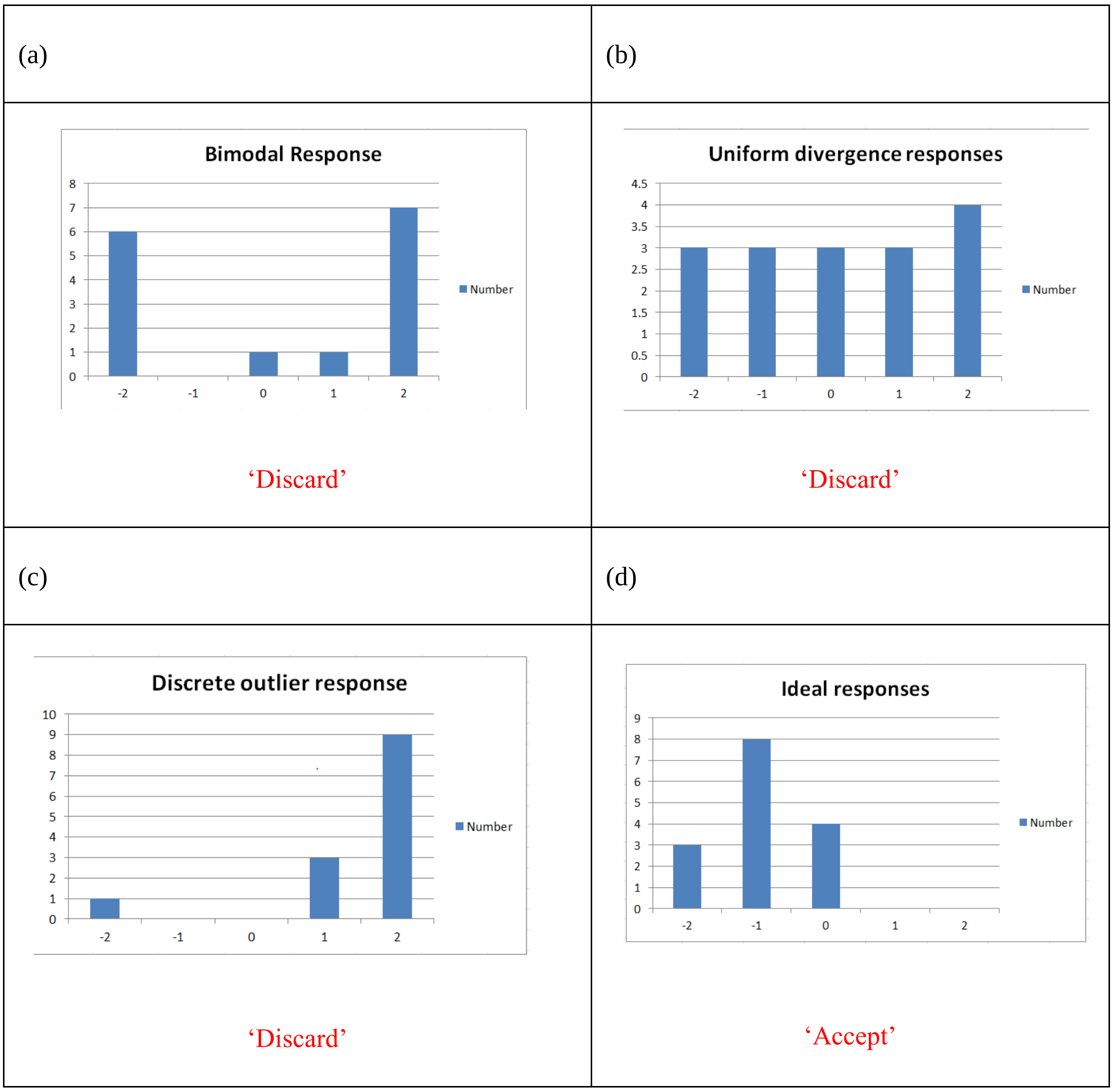
Test optimisation processes

After each panel's scoring, a test optimisation process is conducted where questions with (a) bi-modal, (b) uniform divergence and (c) discrete outlier responses from the panel are discarded, reducing expert disagreement in the answers. The remaining items are all with the (d) logical 'ideal' response pattern from the expert reference panel, to ensure accuracy of content in the SCT items. As a result, only items when the expert panel mostly agrees about the correct responses (the ‘ideal’ response) are selected for the SCT examination. Examples of these responses are shown in Figure 1 (2).

This optimisation process is an existing inherent validity measure in the SCT development process, and is quite different from the usual ‘canonical’ approach in SCT item selection (25). In each year, as a result of this optimisation process, around 20-30% of the SCT items are discarded or modified because of this discordance in response pattern amongst clinicians (i.e. extreme expert disagreement) in the panel. The process is an important quality control measure in SCT examination development to ensure both the content and construct validity of the test.

Figure 1. Expert panel responses to questions in script concordance test.

(a) Bi-modal response, (b) Uniform divergence response, (c) Discrete outlier response and (d) Ideal response.



Starting from 2013, apart from fulfilling the usual assessment blueprint and the above mentioned test optimisation process, an additional quality assurance process has been in place to ensure each SCT paper is made up of items with roughly equal distribution of extreme ('-2' or '+2') and median ('-1', '0' or '+1') modal responses by the expert reference panel. This is referred to here as the 'balanced approach'.

A sample SCT examination with this 'balanced approach' is shown in Table 3.

Table 3. Sample 2016 SCT examination showing Panelist's scores and spread of full marks (modal responses) across median and extreme responses (the 'balanced approach').

2016 Year 4 Q no.	A (-2)	B (-1)	C (0)	D (+1)	E (+2)
1	0.00	0.00	0.00	0.75	1.00
2	0.00	0.56	1.00	0.00	0.00
3	0.44	1.00	0.11	0.00	0.00
4	0.00	0.00	0.00	0.75	1.00
5	0.00	0.00	0.00	0.44	1.00
6	0.00	0.08	1.00	0.00	0.00
…	…	…	…	…	…
37	1.00	0.29	0.00	0.00	0.00
38	0.00	0.00	0.00	0.80	1.00
39	0.00	0.00	0.17	1.00	0.67
40	0.33	0.50	1.00	0.00	0.00
Spread	**8**	**9**	**6**	**5**	**12**

	Total
Extreme	**20**
Median	**20**

Data Analysis

Simulation 1

Simulation 1 involved post-hoc recoding of all '-2' responses to '-1' and '+2' responses to '+1' respectively. Actual average SCT scores and the average scores after simulated rescoring, for the whole cohort were firstly analysed by score quartiles, to investigate a possible ability-treatment effect. Then, average SCT scores for the bottom 10% in each academic year from 2012–2016 were compared with the respective simulated average scores using the paired sample *t*-test.

Simulation 2

The second simulation involved scoring as if all students had chosen '0', the 'neutral' response option in the middle of the Likert style response scale, for their responses in all 40 SCT items.

An example of the two simulations is shown in Table 4. The two simulations were performed to replicate the two previous studies raising concerns about the validity threats to SCT scores as a result of a potential student strategy of avoiding extreme answer options (17, 23). To also investigate the impact of a ‘balanced approach’ to SCT test construction which has been adopted since 2013, pre-intervention scores from 2012 were compared to post-intervention scores (2013-2016). The paired sample *t*-test was used for statistical analysis of the comparisons (IBM SPSS 24).

Table 4: Example of recoding involved in simulations 1 and 2.

SCT Item No.	Actual Response	Simulation (1)	Simulation (2)
1	-2	-1	0
2	0	0	0
3	+1	+1	0
4	+2	+1	0
5	-2	-1	0
6	-1	-1	0

Ethics approval was obtained from the University’s Human Research and Ethics Committee.

Results:

Distribution of modal responses from the expert reference panel: SCT 2012 - 2016

From 2012 to 2016, SCT has been introduced as part of the assessment program for the Bachelor of Medicine, Bachelor of Surgery (MBBS) course and a total of 10 cohorts of 120 students each have been examined in this time. Since 2013, with the ‘balanced approach’, there was a balance of items with extreme (‘-2’ or ‘+2’) modal responses (45-55%) and median (‘-1’, ‘0’ or ‘+1’) modal responses (45-55%). The actual distribution of the median and extreme responses among the 40 items examination for the 5 years is shown in Table 5.

Table 5. Distribution of the modal responses from expert reference panels and the average credit point for each answer response option in 10 SCT papers (2012-2016).

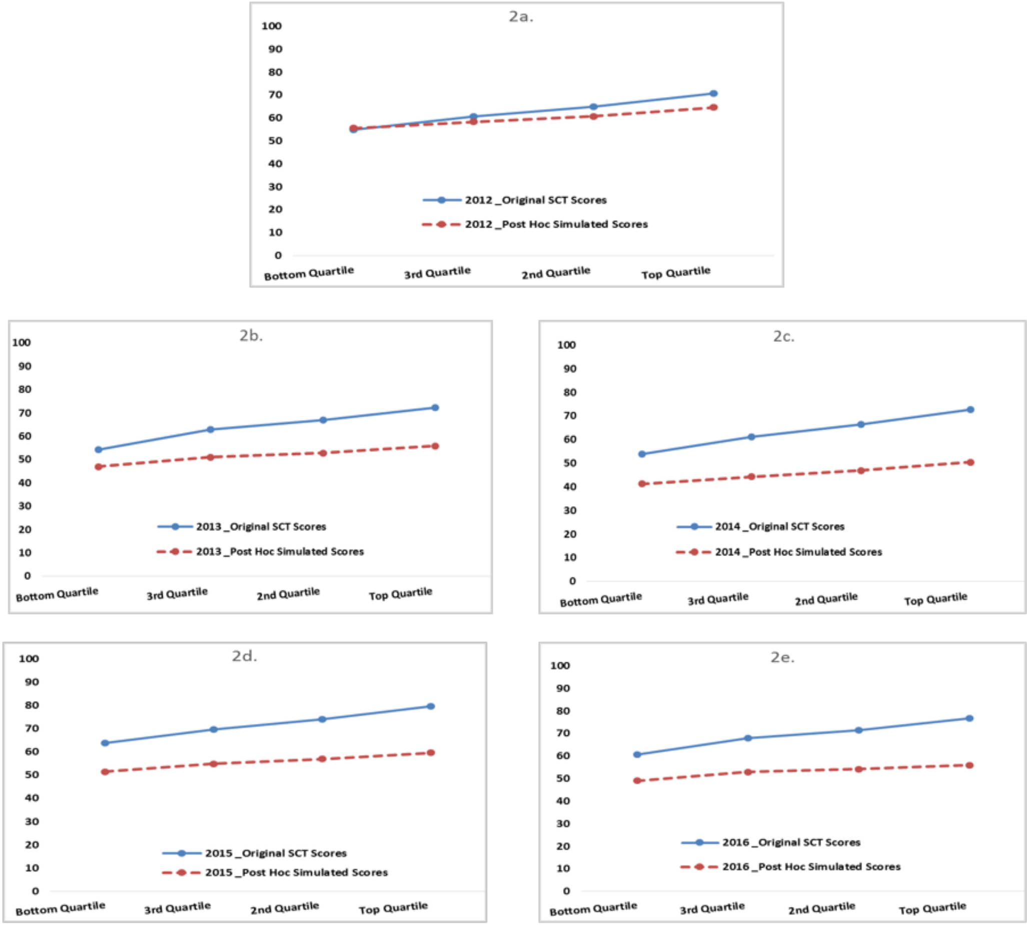
	Number of SCT items with modal answer: a b c d e (Average credit point for each response options)					Total Number of SCT items with median options as modal answers (b + c + d)	Total Number of SCT items with extreme options as modal answers (a + e)	Ratio of median options as modal answer to extreme options as modal answer
	-2	-1	0	+1	+2			
2012 Year 3*	11 (0.38)	11 (0.46)	4 (0.24)	10 (0.31)	4 (0.16)	25	15	63:37
2012 Year 4*	11 (0.43)	9 (0.49)	12 (0.51)	6 (0.24)	2 (0.13)	27	13	68:32
2013 Year 3	15 (0.47)	9 (0.38)	2 (0.14)	9 (0.28)	5 (0.15)	20	20	50:50
2013 Year 4	12 (0.40)	6 (0.25)	8 (0.26)	8 (0.3)	6 (0.22)	22	18	55:45
2014 Year 3	10 (0.35)	7 (0.27)	7 (0.24)	6 (0.25)	10 (0.25)	20	20	50:50
2014 Year 4	13 (0.37)	5 (0.24)	8 (0.28)	7 (0.26)	7 (0.28)	20	20	50:50
2015 Year 3	12 (0.42)	9 (0.32)	5 (0.22)	4 (0.24)	10 (0.31)	18	22	45:55
2015 Year 4	9 (0.30)	8 (0.33)	8 (0.35)	5 (0.29)	10 (0.28)	21	19	52:48
2016 Year 3	12 (0.46)	11 (0.37)	4 (0.18)	3 (0.19)	10 (0.29)	18	22	45:55
2016 Year 4	9 (0.30)	7 (0.33)	6 (0.25)	5 (0.30)	13 (0.39)	18	22	45:55

* Before using the ‘balanced approach’ in the examination

Simulation 1 - Effect on SCT Scores

The effect of simulation 1 (recoding of extreme answer options) on SCT scores for all students in each academic year, analysed by quartile, was examined. Figure 2 (2a-e) shows the original and simulated scores from 2012-2016 respectively, by quartiles. In 2012 (Figure 2a), before the 'balanced approach' was introduced, the avoidance of extreme answer options, did not seem to have huge impact on SCT scores, as indicated by the closeness between the two line graphs for original and post-hoc simulated recoded scores. Figure 2a also shows that in 2012, students in the bottom quartile may have scored slightly higher SCT scores simply by avoiding the extreme answer options. For 2013-2016 (with the 'balanced approach), avoidance of extreme answer options has clearly resulted in a significantly lower score across the whole cohort irrespective of the performance quartile.

Figure 2. Original and simulated scores from 2012-2016 according to the performance ability in quartiles of the students.



Due to our experience and reports from the literature concerning the test taking strategies potentially adopted by poorer performing students, the actual SCT scores and the post-hoc recoded scores of the bottom 10% of the students in each cohort in the 10 SCT papers were also compared using the paired sample *t*-test. Table 6 shows the average SCT scores of the Year 3 and Year 4 cohorts in 2012, as well as, the average SCT scores of the 8 cohorts of Year 3 and Year 4 students from 2013 to 2016. At baseline in 2012 (i.e. before the implementation of the ‘balanced approach’), comparison of the individual cohort SCT exam data revealed that in the Year 3 SCT paper, there was a statistically significant higher average SCT score (55.59%) in simulation 1 (post-hoc rescoring to simulate avoidance of extreme answer options) compared to the original average SCT score (50.15%) for the bottom 10% of the cohort. In contrast, for Year 3 and 4 medical students from 2013 to 2016, post-hoc recoding of extreme response options into median response options resulted in statistically significant lower average SCT scores.

Table 6. Effect of deliberate avoidance of extreme answer options on SCT scores – An investigation through post-hoc simulated rescoring of responses from bottom 10% of students in each SCT paper.

	Mean SCT Scores After Post-Hoc Rescoring (SD)	Original Mean SCT Scores (SD)	Mean Difference (SD)	95% Confidence Interval Mean Difference		Statistical Significance of Mean Difference (p-value)
				Lower	Upper	
2012 Year 3#	55.59 *(4.10)*	50.15 *(1.55)*	5.44 *(3.74)*	2.92	7.95	t (10)= 4.818; p =0.001*
2012 Year 4#	52.26 *(5.42)*	53.37 *(4.44)*	-1.10 *(5.22)*	-4.61	2.40	t (10)= -0.702; p =0.50
2013 Year 3	48.66 *(3.08)*	49.53 *(3.07)*	-0.88 *(2.63)*	-2.65	0.89	t (10)= -1.102; p =0.296
2013 Year 4	44.28 *(4.18)*	52.13 *(2.58)*	-7.85 *(3.65)*	-10.30	-5.40	t (10)= -7.138; p < 0.001**
2014 Year 3	37.36 *(5.11)*	46.79 *(5.20)*	-9.43 *(3.59)*	-11.84	-7.02	t (10)= -8.714; p < 0.001**
2014 Year 4	42.68 *(2.39)*	52.57 *(2.46)*	-9.89 *(2.70)*	-11.55	-8.22	t (10)= -13.243; p < 0.001**
2015 Year 3	45.03 *(2.81)*	58.59 *(2.72)*	-13.55 *(2.87)*	-15.38	-11.73	t (11)= -16.36; p < 0.001**
2015 Year 4	54.29 *(4.30)*	63.16 *(3.50)*	-8.87 *(3.20)*	-11.03	-6.72	t (10)= -14.255; p < 0.001**
2016 Year 3	42.96 *(3.53)*	56.43 *(2.79)*	-13.47 *(3.27)*	-15.55	-11.39	t (11)= -18.121; p < 0.001**
2016 Year 4	51.85 *(4.14)*	57.24 *(3.51)*	-5.39 *(4.15)*	-8.03	-2.745	t (11)= -4.494; p < 0.001**

# 2012: Before the implementation of the 'balanced approach'

* *Statistically significant (at p< 0.05) mean difference between original scores and scores after post-hoc rescoring to simulate deliberate avoidance of extreme answer options*

** *Statistically significant (at p< 0.001) mean difference between original scores and scores after post-hoc rescoring to simulate deliberate avoidance of extreme answer options*

These comparisons are represented pictorially in Appendix A (online) to highlight the findings: for the bottom 10% of students in the 2012 Year 3 cohort, avoidance of extreme answer options has resulted in significant SCT score increase; while in 2013 to 2016 after the adoption of the balanced approach, the same answering strategy produced significantly lower overall SCT scores for the bottom 10% of students in each paper.

Simulation 2 - Effect on SCT Scores

In simulation (2), the data for the 2012 Year 4 cohort, indicates that a student could theoretically score a pass (51.3%) just by choosing ‘0’ to all questions in the examination. In 2013 to 2016, after the implementation of the ‘balanced approach’ in compiling each SCT paper, the same test-taking behaviour would result in a definitive fail in the examination with a score from 13.7% - 35% (Table 7).

Table 7: Simulated scores if students chose ‘0’ for all items compared to actual cohort mean scores in Year 3 & 4 from 2012-2016 (simulation 2).

Year	Year 3 Simulated Scoring (%)	Year 3 SCT Actual Cohort Mean Score (%)	Year 4 Simulated Scoring (%)	Year 4 SCT Actual Cohort Mean Score (%)
2012*	**38.8**	60.5	51.3	65.2
2013	**13.7**	62.0	**25.7**	66.1
2014	**24.5**	70.0	**27.5**	64.6
2015	**35.0**	62.5	**22.0**	73.5
2016	**18.0**	68.3	**24.8**	70.4
Average 2013-16	**22.8**	65.7	**25.0**	68.7

*2012: Before the use of the ‘balanced approach’ in the examination

All failed scores highlighted in bold

Discussion

This study has investigated a possible intervention that may mitigate one of the threats to SCT validity due to construct irrelevant differences in examinee's response style. The latter is an issue previously raised by authors such as Lineberry (17) who had shown that a medical student's use of the strategy of avoiding extreme answer options in SCTs may potentially impact on the validity of his/her test results. The current study has shown that SCT test optimisation processes such as balancing the distribution of expert reference panel's modal answers for items in SCT, across the whole continuum of the Likert response scale, and controlling for other conceptual/logical flaws in partial aggregate scoring used for the conventional SCT, has the potential to further enhance the validity of SCT scores. More specifically, with the addition of the 'balanced approach' as an additional step in the test optimisation processes for SCT, examinees who potentially choose to deliberately avoid extreme answer options, and/or, simply select the 'neutral' answer options, would get significantly lower, rather than higher SCT scores.

In both simulation 1 and simulation 2, using baseline data from 2012 for the 10% of students with the lowest SCT scores (before the adoption of a 'balanced approach'), deliberate avoidance of extreme answer options seemed to result in signs of score 'inflation' (i.e. increase in SCT scores). This is consistent with the findings reported from previous studies (17, 23). The analysis of baseline data for all students in 2012 prior to the implementation of 'balanced approach' also shows that there was an interaction effect between examinees' ability (using scores quartile as proxy) and the prevalence and extent of score 'inflation' through deliberate avoidance of extreme answer options (Figure 2a). There is supporting evidence for this from a recent study that demonstrated that students whose SCT scores are in the lowest quartile are more likely to use this test-taking strategy (avoidance of extreme response options) (24). The current study has provided further empirical evidence that the implementation of the 'balanced approach' since 2013 has mitigated the concern that a test-taking strategy could result in an increase in SCT scores, and threaten the validity of the SCT scores.

In fact, with the balancing of the SCT items in the examination, the potential strategic answering approach would result in a much lower score, as demonstrated in the two simulations conducted in this study. We therefore suggest that careful construction of SCT items and an additional test optimisation process based on expert reference panel's response pattern in SCT items (the 'balanced approach'), could remove the potential "score inflation" previously described (17).

Improving the validity of Script Concordance Testing

When informing students about the structure of SCT and how to appropriately approach and answer SCT items, it may be necessary and beneficial to emphasise the fact that, as an inherent validity feature built-in the design of each SCT paper, there is always a somewhat balanced distribution of median and extreme options in expert panel's modal answers which will attract a full mark. Students should be urged to respond according to their knowledge, understanding and reasoning of all available information for each case. Any deliberate attempt to use any test taking strategy will not be advantageous, but, on the contrary, may disadvantage their SCT scores.

Apart from fulfilling the usual assessment blueprint requirements, a balanced distribution of extreme and median options in the modal responses by expert reference panel, is certainly a useful validity feature that can be built in the SCT test development process, i.e. as the final step in the routine SCT test optimisation process. This is particularly important if a partial credit aggregate scoring algorithm is used.

It is important to acknowledge that there are other validity concerns about SCT, particularly in relation to the logical inconsistency of the answer responses and the accuracy of the expert panel answers compared to the evidence based likelihood ratios

(16, 17). The pre-existing test optimisation procedures adopted in the study context have addressed the concern about the faulty logic of aggregated scoring in SCT (17). The selection of only the 'ideal' responses from the panel and reducing the expert disagreement results in a test focusing on clinical reasoning and data interpretation for clinical scenarios with relatively clear modal answers. Examinees will still score if they veer slightly in either direction from the modal answer response. Elimination of items with 'bi-modal' and 'discrete outliers' response pattern from the expert reference panel, through the pre-existing test optimisation procedures, has somewhat alleviated the extreme complications in reliability estimation from the usual canonical SCT aggregated score.

This study has only been able to focus on addressing one of the validity concerns re SCTs. Others such as SCT standard setting, are issues for further investigation elsewhere. Limitations of the current study also include the unequal number of data points for pre (2012) and post intervention with a balanced approach (2013–2016); and the fact that this is a study in the context of one medical school. The former arose due to the need to adopt the balanced approach to SCT item construction and test optimisation, once the potential threat of the aggregate partial credit scoring methods to test validity was revealed. The potential for score inflation by low performing students avoiding extreme response options was a result found in our, as well as other, settings.

We need to ensure that the items selected in the rigorous item optimisation procedures, do not lead to deviation from the assessment blueprint established for each SCT paper. In other words, the content validity– particularly its alignment with the construct of interest, i.e. decision making in the context of clinical uncertainties in the real clinical setting, should not be compromised as a result of deliberate measures to mitigate the potential for a test-taking strategy to increase student scores and threaten SCT validity. Sharing of a larger pool of SCT items spreading across disciplines with other medical schools would facilitate development of a database of carefully constructed and high content validity SCT items aligned with the construct of interest for use in the assessment of all senior medical students.

To further understand the utility of SCT in assessment of undergraduates' clinical reasoning, the think-aloud method has been proposed to allow the students to justify the reasons for choosing a particular response option in answering the SCT (26). This process may help address the concern raised by Kreiter (2012) that there is no firm evidence of the clear relationship between the purported construct of the SCT (clinical data interpretation) and the response process of examinees (27). Indeed in a response to this, Lubarsky et al have suggested 'think-aloud' or concept mapping protocols might also help to shed further light on examinees' use of probability versus typicality-based reasoning strategies in responding to SCT items (28). As a result of the current study, we highly recommend investigation and routine monitoring of evidence for possible validity threats in SCT scores when SCTs are used for summative purposes.

Conclusion

We would like to reiterate that, in interpreting the findings from this study, one should note the fact that this simulated investigation on plausible validity threat to SCT scores due to test-wise examinees deliberately avoiding the extreme answer options, was carried out in a context where there has been considerable pre-existing and inherent validity measures in place to control for more fundamental conceptual flaws associated with the aggregate partial credit scoring approach (based on an expert reference panel's responses). It was never the intention to paint a simplistic and reductionistic view through this manuscript, that the 'balanced approach' – i.e. the intervention investigated in this study is the ultimate solution for all potential validity threats and issues with SCT. On the contrary, we recommend that the hypotheses and conclusion derived from this study to be further tested in other medical education settings.

Improving the validity of Script Concordance Testing

References

1. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance Test: A Tool to Assess the Reflective Clinician. Teaching and Learning in Medicine. 2000;12(4):189-95.
2. Wan SH. Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine. HONG KONG MEDICAL JOURNAL. 2015;21(5):455-61.
3. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: From theory to practice: AMEE Guide No. 75. Medical Teacher. 2013;35(3):184-93.
4. Gagnon R, Charlin B, Coletti M, Sauvé E, Van Der Vleuten C. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? Medical Education. 2005;39(3):284-91.
5. Carrière B. Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a Script Concordance Test. Ann Emerg Med. 2009;53(5):647-52.
6. Drolet P. Assessing clinical reasoning in anesthesiology: Making the case for the Script Concordance Test. ANAESTHESIA CRITICAL CARE & PAIN MEDICINE. 2015;34(1):5-7.
7. Duggan P, Charlin B. Summative assessment of 5th year medical students' clinical reasoning by script concordance test: requirements and challenges. BMC MEDICAL EDUCATION. 2012;12(1):29-.
8. Kazour F, Richa S, Zoghbi M, El-Hage W, Haddad FG. Using the Script Concordance Test to Evaluate Clinical Reasoning Skills in Psychiatry. Academic Psychiatry. 2017;41(1):86-90.
9. Nouh T, Boutros M, Gagnon R, Reid S, Leslie K, Pace D, et al. The script concordance test as a measure of clinical reasoning: A national validation study. American Journal of Surgery. 2012;203(4):530-4.
10. Tsai T-C, Chen D-F, Lei S-M. The ethics script concordance test in assessing ethical reasoning: really good stuff. Medical Education. 2012;46(5):527-.
11. Faucher C, Dufour‐Guindon MP, Lapointe G, Gagnon R, Charlin B. Assessing clinical reasoning in optometry using the script concordance test. Clinical and Experimental Optometry. 2016;99(3):280-6.
12. Dumas JP, Blais JG, Charlin B. Script concordance test: can it be used to assess clinical reasoning of physiotherapy student? Physiotherapy. 2015;101:e332-e3.
13. Lubarsky S, Vleuten CPMvd, Charlin B, Chalk C, Cook DA. Script concordance testing: a review of published validity evidence. Medical Education. 2011;45(4):329-38.

14. See KC, Tan KL, Lim TK. The script concordance test for clinical reasoning: re-examining its utility and potential weakness. Medical Education. 2014;48(11):1069-77.

15. Wan SH. Using Script Concordance Testing (SCT) to Assess Clinical Reasoning-The Progression from Novice to Practising General Practitioner. Medical Education. 2014;48(2):6.

16. Ahmadi S-F, Khoshkish S, Soltani-Arabshahi K, Hafezi-Moghadam P, Zahmatkesh G, Heidari P, et al. Challenging script concordance test reference standard by evidence: do judgments by emergency medicine consultants agree with likelihood ratios? International Journal of Emergency Medicine 2014;7.

17. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Medical Education. 2013;47(12):1175-83.

18. Ducos G, Lejus C, Sztark F, Nathan N, Fourcade O, Tack I, et al. The Script Concordance Test in anesthesiology: Validation of a new tool for assessing clinical reasoning. ANAESTHESIA CRITICAL CARE & PAIN MEDICINE. 2015;34(1):11-5.

19. Erickson G, Wagner K, Morgan M, Hepps J, Gorman G, Rouse C. Assessment of Clinical Reasoning in an Environment of Uncertainty: A Script Concordance Test for Neonatal-Perinatal Medicine. Academic Pediatrics. 2016;16(6):e6.

20. Humbert AJ, Miech EJ. Measuring Gains in the Clinical Reasoning of Medical Students: Longitudinal Results From a School-Wide Script Concordance Test. Academic Medicine. 2014;89(7):1046-50.

21. Cullen MJ, Sackett PR, Lievens F. Threats to the Operational Use of Situational Judgment Tests in the College Admission Process. International Journal of Selection and Assessment. 2006;14(2):142-55.

22. McDaniel MA, Psotka J, Legree PJ, Yost AP, Weekley JA. Toward an understanding of situational judgment item validity and group differences. The Journal of applied psychology. 2011;96(2):327-36.

23. See KC, Tan KL, Lim TK. The script concordance test for clinical reasoning: re‐examining its utility and potential weakness. Medical Education. 2014;48(11):1069-77.

24. Wan SH, Duggan P, Tor E, Hudson JN. Association between candidate total scores and response pattern in script concordance testing of medical students. Focus on Health Professional Education: A Multi-disciplinary Journal. 2017;18(2):26-35.

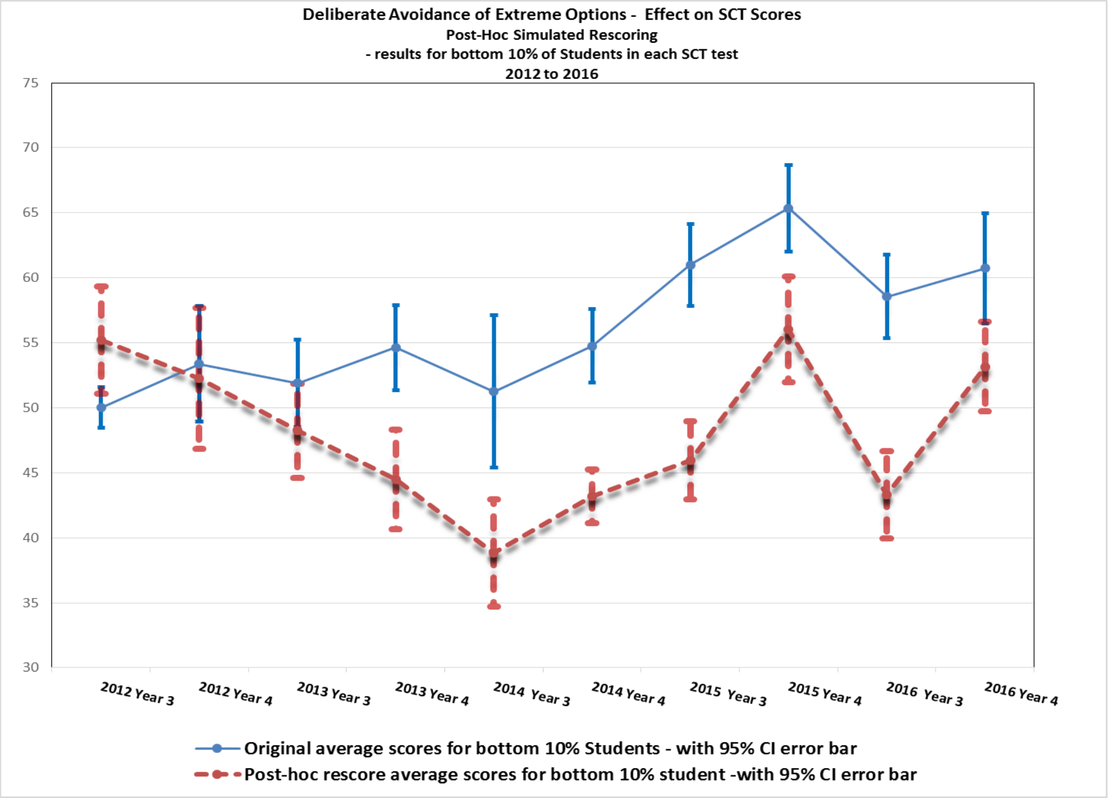
25. Fournier JP, Demeester A, Charlin B. Script concordance tests: Guidelines for construction. BMC Medical Informatics and Decision Making. 2008;8(1):18-.

26. Power A, Lemay J-F, Cooke S. Justify Your Answer: The Role of Written Think Aloud in Script Concordance Testing. Teaching and Learning in Medicine.

2017;29(1):59-67.
27. Kreiter CD. Commentary: The response process validity of a script concordance test item. Advances in Health Sciences Education. 2012;17(1):7-9.
28. Lubarsky S, Gagnon R, Charlin B. Script concordance test item response process: The argument for probability versus typicality. Advances in Health Sciences Education. 2012;17(1):11-3.

Online attachment:

Appendix A. Effect on SCT scores by deliberate avoidance of extreme options for bottom 10% of students (2012-2016).


Deliberate Avoidance of Extreme Options - Effect on SCT Scores
Post-Hoc Simulated Rescoring
- results for bottom 10% of Students in each SCT test
2012 to 2016
2012 Year 3
2012 Year 4
2013 Year 3
2013 Year 4
2014 Year 3
2014 Year 4
2015 Year 3
2015 Year 4
2016 Year 3
2016 Year 4
Original average scores for bottom 10% Students - with 95% CI error bar
Post-hoc rescore average scores for bottom 10% student -with 95% CI error bar