2020

Exploring the Validity of Script Concordance Testing to Assess the Clinical Reasoning of Medical Students

Micahel SH Wan
The University of Notre Dame Australia

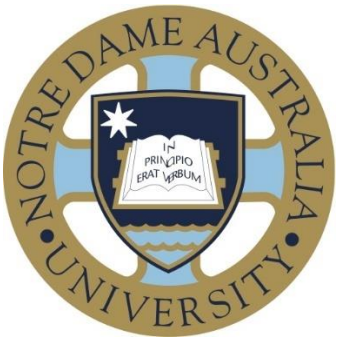
EXPLORING THE VALIDITY OF SCRIPT CONCORDANCE TESTING TO ASSESS THE CLINICAL REASONING OF MEDICAL STUDENTS

Michael Siu Hong WAN

MBChB, MRCP (UK), FRCP (Edin), FHKCP, FHKAM, GCUT, FANZAHPE

Submitted in fulfilment of the requirements for the

Doctor of Philosophy (by publication)



School of Medicine

Sydney Campus

The University of Notre Dame Australia

May, 2020

Declaration of authorship

To the best of the candidate's knowledge, this thesis contains no material previously published by another person, except where due acknowledgement has been made.

This thesis is the candidate's own work and contains no material which has been accepted for the award of any other degree or diploma in any institution.

The research presented and reported in this thesis was conducted in accordance with the National Health and Medical Research Council National Statement on Ethical Conduct in Human Research (2007, updated 2018). The proposed research study received human research ethics approval from the University of Notre Dame Australia Human Research Ethics Committee (EC00418), Approval Number # HREC 016126S, 018161S, 019023S.

Signed:

Date: 25/05/2020

Abstract

Assessment of clinical reasoning is often challenging, as it is a complex process of thinking and decision making. Script Concordance Testing (SCT), using authentic clinical scenarios with diagnostic or management uncertainties, has been developed to assess clinical reasoning. As SCT is a relatively new assessment modality, more empirical evidence is needed to support the validity of SCT scores.

This thesis examines key aspects of the validity of SCT scores in the assessment of the clinical reasoning ability of medical undergraduates. A review of the current literature informs the use, design and standard setting of SCT, as well as evidence for its reliability and validity. Exploration of the response patterns of 5 cohorts of graduate-entry medical students in an Australian Medical School showed deliberate avoidance of extreme responses by the lowest quartile students. A post-hoc simulation study, testing the hypothesis that test-wise candidates' SCT scores were inflated through deliberate avoidance of extreme response-options and selection of neutral response-options, generated an approach to optimising and balancing SCT items for improved SCT score validity.

In response to the paucity of empirical studies on the construct validity for SCT scores, the next study showed evidence of progression in SCT scores from medical students, to junior registrars, to experienced general practitioners. Finally, an investigation of candidates' response process, using a 'think-aloud' approach, supported the response process validity of SCT scores.

In conclusion, this thesis has demonstrated that: 1) thoughtful design and balance of SCT items can mitigate some of the validity threats to medical student SCT scores; 2) the tendency of SCT scores to progress with increasing levels of clinical practice experience further supports the construct validity of SCT scores; and 3) use of the 'think-aloud' approach to explore students' response process may enhance the utility and educational benefits of SCT. The research supports the validity of SCT in assessing clinical reasoning in undergraduate medical education, and presents practical approaches to enhance the design of the assessment instrument.

Acknowledgements

I would like to express my greatest gratitude to my two supervisors Associate Professor Elina Tor and Professor Nicky Hudson for their patience and guidance throughout the years. Together, they have provided me with encouragement and expert advice on the research project. They have inspired me to take up new challenges, to focus on the relevant areas and their mentoring helped me to think outside the box. I now have a better understanding of the literature on Script Concordance Testing. I am extremely fortunate to have them as my supervisors of this PhD research and I greatly appreciate their dedication and commitment on top of their ongoing busy academic work.

I would like to also acknowledge the Australian Government for funding the project under the RTP scheme.

None of this research would be possible without the unconditional support of my wife Irene and my daughter Beatrice. Their patience and tolerance during my doctorate research have given me the encouragement to persevere. I am grateful for their warm reassurances during my stressful times.

Professor John Wong, my former teacher in Pharmacology during my undergraduate medical training days, has inspired me to take up the PhD endeavour and his continual spiritual support and guidance have been invaluable.

I would also like to thank Professor George Mendz (Head of Research) for his guidance and advice for the thesis; Professor Rufus Clarke for his support in my early research development; and Miss Eunice Lau of the assessment team in the School for her help in the data capturing and technical advice in the formatting of the manuscripts.

List of Publications

1. **Wan M.** Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine. Hong Kong Medical Journal (**ERA Journal**). 2015;21(5):455-61.
2. **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 (**ERA Journal**). 2017;18(2):26-35.
3. **Wan MS**, Tor E, Hudson JN. Improving the validity of script concordance testing by optimising and balancing items. Medical Education (**ERA Journal**). 2018;52(3):336-46.
4. **Wan SH**, Tor E, Hudson JN. Construct validity of script concordance testing: progression of scores from novices to experienced clinicians. International Journal of Medical Education (IJME) (**ERA Journal**). 2019;10:174-9.
5. **Wan SH**, Tor E, Hudson JN. Commentary: expert responses in script concordance tests: a response process validity investigation. Medical Education (**ERA Journal**). 2019;53(7):644-6.
6. **Wan SH**, Tor E, Hudson JN. Examining response process validity of Script Concordance Testing: a think-aloud approach. International Journal of Medical Education (IJME) (**ERA Journal**). 2020;11:127-135.

Table of Contents

Chapter 1: Introduction

1.1 What is Clinical Reasoning and why is it important in health professional education?

As defined very succinctly by Higgs, '*Clinical reasoning is the sum of the thinking and decision-making processes associated with clinical practice'*. (1) It is the process by which the clinician collects and processes the patient's information in order to understand the underlying problem, plans and implements the management, followed by evaluation and reflections on the outcomes. (2) The graphical representation of this reasoning cycle is shown in Figure 1. A sound knowledge base of basic and clinical sciences, together with an understanding of disease pathophysiology, are required for the effective learning and application of clinical reasoning.

In the clinical setting, a healthcare professional usually starts with gathering a comprehensive history of the presenting illness from the patient, followed by a focused physical examination to collect all the relevant information. The resulting symptoms and signs are then analysed to allow formulation and generation of the hypotheses. By recalling the relevant knowledge and illness scripts in his or her mind, assisted by previous experiences in similar presentations, the health professional will formulate an appropriate investigation and management/action plan for the patient's problems/issues. This clinical judgement will take into account the cultural background, social and psychological aspects of the patient presentation, and the context of the health system, institution and the state or country. This will be followed by an evaluation of the outcomes (physical, psychological and social) and a self-reflective process that consolidates the learning to aid future improvement in patient management. (3) The reasoning cycle generates a range of illness scripts in a developing clinicians' mind as they gain experience. This process is very efficient as the script activation is automatic and almost unconscious. (4)

The terms clinical reasoning and clinical judgement are often used interchangeably in the literature. However, one should be mindful that clinical reasoning is a process, whereas clinical judgement is the result of the process and represents the decisions that a clinician makes. (5)

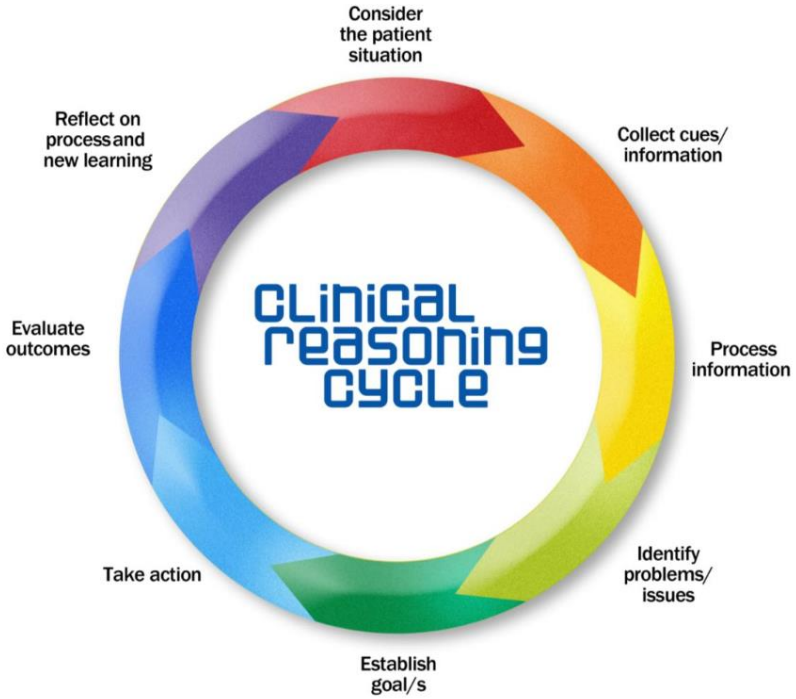


Figure 1: Clinical Reasoning Cycle. Retrieved from: http://www.utas.edu.au/__data/assets/pdf_file/0003/263487/Clinical-Reasoning-Instructor-Resources.pdf. Clinical reasoning instructor resources. School of Nursing and Midwifery. Faculty of Health, University of Newcastle. 2009.

Clinical reasoning is a foundation requirement for entry level clinical practice and cannot be assumed to develop in the absence of fundamental educational strategies. Effective clinical reasoning skills have a positive impact on patient outcomes, minimising adverse medical events and reducing diagnostic errors. (3, 6, 7) Competent clinicians must be able to apply correct clinical reasoning in day to day practice for efficient and safe patient-centred care. A significant number of patient problems are managed inappropriately due to diagnostic and management errors. (8, 9) Improving critical thinking (the process of intentional higher level thinking) and clinical reasoning skills of medical students as they journey through medical school to

fellowship training is essential to ensure optimal patient outcomes. Sound and effective clinical reasoning skills are essential for practising clinicians.

There are many cognitive processes that underpin the construct of clinical reasoning. Theories of the cognitive processes that underpin clinical reasoning have been described as 'normative' or 'descriptive'. (10) For the 'normative' theories, Kassirer described three types of reasoning namely: probabilistic; casual and deterministic. (11) Probabilistic reasoning relies on Bayes' theorem, a simple mathematical formula used to calculate conditional probabilities. Disease prevalence, sensitivity and specificity of diagnostic tests are taken into account to support the diagnosis. However, this approach is limited by the fact that sometimes it is difficult to describe a test to be either just positive or negative; and the Bayesian analysis requires that each disease is mutually exclusive of all other disease (which is highly unlikely in the actual real world clinical setting). Casual reasoning however uses the underlying basic sciences principles (e.g. anatomy, physiology, biochemistry) and pathophysiology to explain and guide treatment. Casual reasoning is therefore considered as a foundation to other reasoning models. Deterministic reasoning uses branching chain style algorithms in which the logic can be clearly defined to guide diagnosis, treatment or prognosis. However, the limitation of this logical reasoning is that it fails to deal with the uncertainties in the clinical environment.

For the 'descriptive' theories, there are two main cognitive capabilities and processes that underpin the use of clinical reasoning. The first approach, 'hypothetical deductive reasoning' (also known as the analytical/Type 2) approach uses data collection, hypotheses generation, data interpretation followed by hypothesis evaluation, to arrive at the diagnosis and correct management of a patient. This approach is more commonly used by novices who are usually medical students or junior registrars under training. The second approach is the 'pattern recognition' (also known as the intuitive/Type 1) approach where key symptoms and signs from the patient are perceived as patterns like pieces of a jigsaw puzzle and these are compared to the patterns residing in the memory of the clinician from previous experiences. Once a particular familiar pattern is recognised and formed, a provisional diagnosis is made, with appropriate

investigations and management following. The latter process is much quicker and more efficient and is used by expert clinicians with years of clinical experiences. (9) However, if the expert encounters a clinical scenario where the 'patterns' do not align or match up, he or she will fall back to the first logical step-by-step analytical approach to solve the problem. (12) As these two forms of processing are not mutually exclusive, the expert clinician in real-world clinical settings is able to coordinate both the analytic and nonanalytic (intuitive) cognitive processes to solve and manage a clinical problem (dual processing). (12) While the intuitive approach is expected to dominate in the initial phases of managing a new patient, the analytic approach plays a dominant role during the hypothesis testing stage of the clinical encounter. (13)

As discussed above, there is not a universally accepted model of clinical reasoning. The process of clinical reasoning is not easily captured in any one model. (10) Each model has its own limitations and this does not mean that a particular model is better than the other. During a clinical encounter, most of these models do not consider the patient's social setting (e.g. appearance, language, lifestyle and ways of dressing), non-verbal communications, the doctor-patient relationship and the clinician's personality, feeling or socio-cultural characteristics. (14) A combined or multidimensional model may be better to describe how a clinician actually makes clinical decisions.

As a competent clinician, apart from managing the typical clinical case presentations, one must be capable of dealing with uncertainties and ill-defined problems. (15) Often in clinical encounters with patients, clinical reasoning and decision making have to be made based on incomplete clinical information (from the history taking and/or physical examination) at the initial presentation. This initial uncertainty often results in differences in clinical judgements between expert clinicians. This can present challenges, not only in facilitating clinical reasoning ability in learners, but in assessing the outcomes of learners' clinical reasoning processes. Therefore it can be difficult to define a single correct answer for an assessment item targeting the clinical decision making process and outcomes. A variety of assessment modalities have been developed to assess clinical reasoning competency in medical education. Some will be briefly discussed, before focusing on Script Concordance Testing (SCT), an approach

that was designed to mirror the clinical reasoning process and outcomes in the context of uncertainties in clinical practice.

1.2 Assessment of Clinical Reasoning in medical education

Clinical reasoning has been described as one of the core competencies for medical graduates and fellowship trainees by professional bodies such as the Accreditation Council (ACGME) in the USA, and the Royal College of Physicians and Surgeons in Canada for Graduate Medical Education in Canada. (16, 17)

As explained above, clinical reasoning is a complex process of thinking and decision making, producing challenges in the assessment of the components of this process. The traditional bedside or viva examination, often used to assess clinical reasoning, are usually non-standardised, subjective, often biased and not reliable because of the limited number of examiners and/or cases per student. (18) As a result, a range of assessments has been developed, aiming for greater reliability and validity.

There are currently 7 commonly used assessment tools or formats to assess clinical reasoning in medical education:

- script concordance testing (SCT)
- clinical reasoning problem (CRP)
- key feature problem (KFP)
- comprehensive integrative puzzle (CIP)
- patient management problem (PMP)
- objective structured clinical examination (OSCE)
- workplace based assessments, e.g. the mini-clinical evaluation exercise (mini-CEX)

Some of these tools are briefly discussed before focusing on the use of SCT to assess clinical reasoning, the focus of this thesis.

For the clinical reasoning problem (CRP) format, each problem has a clinical scenario with the patient's clinical presentation, history and physical examination findings. Candidates are asked to nominate the two most likely diagnoses and also asked to explain and list the clinical features that they considered in formulating the diagnoses, indicating whether these features supported or opposed the nominated diagnoses. CRPs (based on the hypothetical deductive model of cognitive processing) were shown in some studies to be both reliable and valid in assessing the accuracy of diagnostic reasoning. (19) However, the marking of CRPs is relatively resource intensive compared to multiple choice question (MCQ) formats such as SCT.

Key feature problems (KFPs) test clinical decision-making skills in written or computer-based formats. After a clinical scenario presentation, candidates are asked to list a few key clinical problems. Then the candidates are asked to choose one or more correct answers from a long list as the appropriate investigation and/or management. KFPs, also based on the hypothetical deductive model, have been used by the Canadian Medical Council and the Royal Australian College of General Practitioners (RACGP) in high stakes examinations with good reliability in combination with other assessment modalities, like MCQs and OSCEs. (20)

The hypothetical deductive cognitive process is also the basis of comprehensive integrative puzzle (CIP) questions, which assess the integrative elements of diagnostic thinking and clinical reasoning. (21) Items in this test are presented in the format of an extended matrix of rows and columns. Candidates are asked to insert the correct information in each cell. The completed horizontal rows reflect integrative ability (diagnostic thinking and clinical reasoning) and the vertical columns measure the student's proficiency in interpreting medical history data, physical examination findings, laboratory test results, and imaging results. (22) Scoring CIP is relatively complex and not well explored and this tool is more suitable for formative assessment or research.

In patient management problem (PMP) items, a clinical scenario is presented followed by sections provided in stages, where candidates need to respond in relation to history

taking, physical examination, investigations and diagnosis. A PMP could be very long and require up to 90 minutes to answer. The low reliability, the problem of case-specificity and lack of evidence showing differentiation of scores between junior and senior doctors have resulted in the low popularity of its use in assessment. (20, 23)

While commonly used written assessments of clinical reasoning (e.g. MCQs, Short Answer Questions – SAQs, and KFPs) have high content validity due to extensive blueprinting, good internal consistency and are relatively easy to be scored, merely selecting the correct answer(s) from a pre-defined list is not truly representative of authentic clinical reasoning. The underlying response process is not evident. (24) The three examination formats (SCT, CRP and CIP) have shown high reliability in a high-stakes national examination designed to test clinical reasoning and decision-making skills in undergraduate medical students in Iran. (25) However there have been limited reports on the use of these three formats to assess clinical reasoning of medical undergraduates.

The OSCE was developed by Harden to be an objective and standardised assessment of clinical competency in medical education. (26) This tool has been used to assess history taking, physical examination, hypothesis generation, management, clinical reasoning, communication skills as well as professionalism. (12) Simulated/ standardised patients, with or without case notes review, can be incorporated into the assessment. However, as it is an integrated assessment of these competencies, the total OSCE score was found to not correlate directly with the clinical reasoning abilities of the medical students. (27) An OSCE can assess medical students' clinical skills including, but not exclusively focusing on clinical reasoning. (24) The OSCE format is also resource intensive and thus one of the most expensive assessment options. (28) Workplace-based assessment (WBA) formats such as Direct Observation, the mini-clinical evaluation exercise (mini-CEX), written notes/report review and global assessment were reported to be useful for assessing clinical reasoning. (24, 29) The OSCE, while aiming for higher reliability, isn't deemed to be as valid as WBA. WBA however is time consuming for an assessor focused on delivering patient care, and can be difficult to stage in a busy workplace.

Recently, customised assessments of clinical reasoning skills in undergraduate medical students such as the Paediatrics Milestone Project targeting key features in the history, examination or laboratory results for medical students and paediatric trainees have emerged, as well as the use of virtual patients in trauma settings. (30, 31) These projects targeted specific sub-specialties or trainee groups with tailored assessment formats.

The SCT tool, on the other hand, can be applied to all disciplines and specialties in the health professional education field, due to its generic format and a layout that can be easily adapted to need. Script concordance testing is a novel modality to assess clinical reasoning and clinical decision making using a response format and scoring based on the Bayesian theory of reasoning. (32) It was designed and developed specifically based on the aforementioned conceptual underpinning of clinical reasoning, mirroring the clinical reasoning process and outcomes in the context of uncertainties in clinical practice. Probabilistic clinical reasoning and decision making are incorporated into SCT through its unique response format, as well as via the scoring approach. The latter uses the aggregated scoring method with a 5-point response scale. The next section elaborates on the introduction of SCT to assess clinical reasoning.

1.3 Using Script Concordance Testing to assess Clinical Reasoning

Acquisition of clinical reasoning skills is a lifelong developmental process across the entire career trajectory of health professionals, particularly for medical doctors. This starts from medical school and continues throughout fellowship training and continual clinical practice. Only recently, has assessment of clinical reasoning been identified as a challenge. While MCQs and SAQs offer the benefit of easy or computer marking, even with ‘well-defined’ problems they generally test at the lower levels of the Miller’s pyramid (Figure 2) and were not considered as ideal for the assessment of clinical reasoning. (33-35)

In 2000, a new method of assessment named Script Concordance Testing (SCT) was introduced by Charlin et al. (36) In SCT, a short real-world authentic clinical vignette is presented in the presence of diagnostic or management uncertainties in an ill-defined setting. This is followed by a question with 3 parts. In Part 1, the provisional diagnosis, investigation or management option is presented. In Part 2, a new piece of clinical information (history symptom, physical sign, laboratory finding or diagnostic imaging finding) is provided. In Part 3, students are asked to indicate the extent to which the emergent new clinical information will affect their clinical decision on the provisional diagnosis; or requesting an investigation or a particular management option. The candidate must choose one of the response options from a 5-point Likert probability scale in the range of '-2' (much less likely/appropriate); '-1' (less likely/appropriate); '0' (neither less, or more likely/appropriate); '+1' (more likely/appropriate); or '+2' (much more likely/appropriate). The candidate's response to each SCT item is compared to that of the expert panel of clinicians who will recall the previous illness scripts in their mind (from their experiences and encounters of similar clinical presentations) to formulate their decision. The closer the candidate's decision to that of the majority of the expert panel members, the higher the score awarded. Hence, the name of 'Script Concordance'. The scoring of SCT items is based on the aggregated scoring system in relation to the probability of diagnosis or action by the expert reference panel. This scoring system is different from classical MCQs where only one best answer will attract a full mark. In SCT, if a response is in concordance with the majority of the expert panel, a full mark will be awarded. A partial weighted score will be awarded for a response that is 'in concordance' with the minority of the panel and a zero mark will be awarded for a non-matching response.

Script concordance testing (SCT) as an assessment tool has the capacity to assess clinical reasoning on ill-defined problems and variability within answers provided by the reference panel. This is a key component of the power of the SCT to discriminate the clinical reasoning ability and performance between different levels of clinical experience. (37)

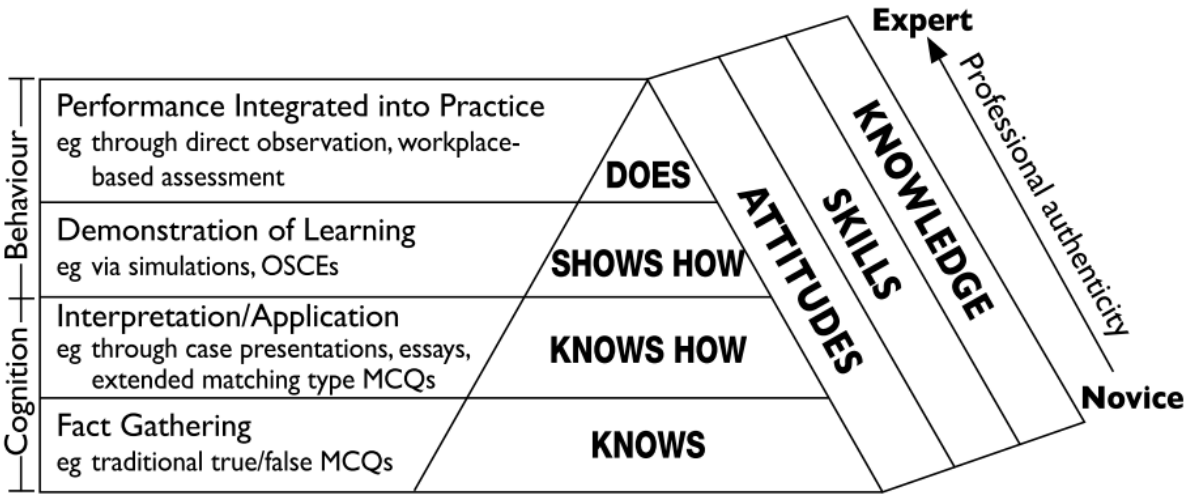


Figure 2. Miller's Pyramid. (38, 39) Reprinted with permission from Dr Ramesh Mehay. Abbreviations: MCQs = multiple-choice questions; OSCEs = objective structured clinical examination. SCT assessing at the 'Knows How' level.

SCT items are written to test the interpretation or application of knowledge (Figure 2) to a selected stage or step in the clinical reasoning process, using the partial aggregate scoring method. SCT mirrors real-life situations by giving students authentic clinical scenarios/vignettes in the question stem, and it can also prioritise defensibility of answer over answer 'correctness'. (28) Another advantage of using SCT to assess clinical reasoning is that the marking/scoring is very similar to classical MCQs with 5 options (A – E), where electronic marking or online examination can be used. Scoring can be done almost instantly using simple formula calculations and no manual marking is required, saving significant time and minimising the workload of academic and administrative staff. (40)

A general narrative review of the literature was conducted at the commencement of this PhD study, which covers SCT design and format; expert reference panel selection; standard setting; issues of reliability and validity in the use of SCT in undergraduate and post-graduate medical education. This is included in Chapter 2 of this dissertation. However the issue of validity of SCT scores, most pertinent for this thesis, is further discussed below in this introduction.

1.4 Validity of SCT Scores

Validity is generally defined as the extent to which an assessment accurately measures what it is intended to measure. According to the *Standards of Educational and Psychological Testing*: (41) 'Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests' (p. 9). In contemporary usage, validity is a unitary concept where construct validity is the whole of validity. Validity always refers to score interpretations and never to the assessment itself. The Standards advise on the five distinct sources of construct validity evidence: content; response process; internal structure; relationship to other variables; and consequences. (42) Sources of evidence for content validity includes the examination blueprint, alignment of the item content to course learning objectives, quality of test questions (whether they are developed according to the accepted best practice) and item developer qualifications. Response process is defined as evidence of data integrity where all sources of error associated with the exam administration are controlled and minimised. Evidence for response process validity also includes quality control on marking, electronic scoring and reporting, accuracy of final marks, the relationship between the intended construct and the underlying thought processes of students, ensuring students understand the examination format and accurate interpretation of student scores. Internal structure, as a source of validity evidence, includes statistical item performance analysis (e.g. item difficulty index and discrimination factor), score reliability, standard errors of measurement (SEM) and other psychometric analyses. Relationship to other variables evidence relates to correlation of assessment scores with other variables such as other modalities of assessment that measure the achievement or ability of the students. Consequential validity evidence refers to the impact of assessment scores on the students or the community. It includes cut score determination, pass-fail consequences, false positives/negatives and the impact of assessments on teaching and learning. (42)

All assessments in medical education require evidence of validity for meaningful interpretation of results. (42) From medical schools to post-graduate fellowship and specialty training, educators have an obligation to the public to ensure the decisions

on the competency of graduates are founded on valid assessment scores. Educators and institutions also need to be able to defend themselves against appeals and challenges to the assessment results, especially in high stakes exit and certification/licensure examinations. Therefore, assessments must be of good quality and be supported by validity evidence. (43)

In designing SCT items, content validity evidence is documented by detailed blueprinting of the questions in relation to the Learning Objectives of the programme and vigorous review of each item by the content experts in the field. For evidence of response process validity it is important to document detailed instructions on how to answer SCTs, ensure quality control processes are in place for data/score accuracy, eliminate poor scoring items and ensure calculations are accurate when using the aggregated scoring system. Similarly, the clear documentation of the cut-score and pass/fail calculations in SCT provide further evidence for consequences validity. (42, 44)

The construct validity of SCT has been reported as depending on the notion that…*candidates with more evolved illness scripts interpret data and make judgments in uncertain situations that increasingly concord with those of experienced clinicians given the same clinical scenarios* (p. 187). (40) Support for the validity of this notion has come from evidence reported in the literature showing that SCT scores consistently increase with increasing level of training. Some examples of this evidence include the progression of medical trainees' scores from Postgraduate Year 1 (PGY1) to PGY3, and to fellows in neonatology as their clinical reasoning skills mature. (45) Progression was also evident across 2 different linguistic, cultural and learning environments in French and Canadian universities. Scores increased with clinical experience in Urology in the two sites. These data further support the stability of the construct validity of the scores from SCT across different learning environments. (46) More evidence of SCT score validity is presented in the literature review paper (Chapter 2) and further explored in the subsequent chapters/ papers of the thesis. The next section addresses some contemporary issues and challenges associated with SCT.

1.5 Current issues and challenges with SCT

As SCT has become more popular and is being implemented in health professional education fields (e.g. in Medicine, Paediatrics, Nursing and Physiotherapy), (47-51) concerns have arisen about the validity of the SCT in assessing clinical reasoning. (52-54) Strictly speaking, validity is not a characteristic of an assessment tool. Validity is a hypothesis on the degree of meaningfulness/appropriateness of assessment scores interpretation and use. Validity needs to be supported with theoretical and empirical evidence from assessment data, and as SCT is a relatively new assessment modality, more empirical evidence is needed to support the validity of SCT scores. The research projects reported in the published manuscripts in this PhD dissertation aimed at gathering more evidence to support (or refute) the validity of SCT scores as measures of the clinical reasoning ability of medical undergraduates. Specifically it addressed the following challenges in the validity of SCT scores: 1) The impact of candidates gaming the SCT examination, by avoiding selection of the extreme response options (i.e. '-2' or '+2'), or deliberately selecting the median ('0') responses, as reported in the literature; (52) 2) The evidence for the construct validity of SCT scores, by exploring the progression in SCT scores from medical students, to junior registrars, to experienced general practitioners; and 3) Investigation of candidates' response process using a written 'think-aloud' approach.

1.6 Aims of the PhD project

The current doctoral project aimed to contribute further evidence for the validity of Script Concordance Testing for assessing the Clinical Reasoning of medical students in the study context. Specifically, this PhD project involved investigations on issues and challenges to the validity of SCT score interpretation, through the following phases of research in a medical school context. It comprised the following:

1. After the initial and general literature review (Chapter 2) looking at the background and use of the SCT in medical education when the thesis research commenced, (55)

the second study and published paper (Chapter 3) investigated the association between SCT scores and response patterns of medical students in the study context. (56) It aimed to provide a background picture for further improvement in the design of SCT papers to mitigate some of the above-mentioned challenges as reported in the SCT literature.

2. The third study and published paper (Chapter 4) in this thesis reported the outcomes of efforts made in the SCT development process to improve the validity of the scores. Specifically, this study investigated the impact of deliberate procedures in balancing the number of SCT items with the extreme and median modal response from the expert reference panel, in addition to the normal test optimisation process, for every SCT paper. This was an initiative in the study context to improve the validity of the SCT scores by minimising the risk of students gaming the exam by avoiding or choosing particular response options in their answers. (57)

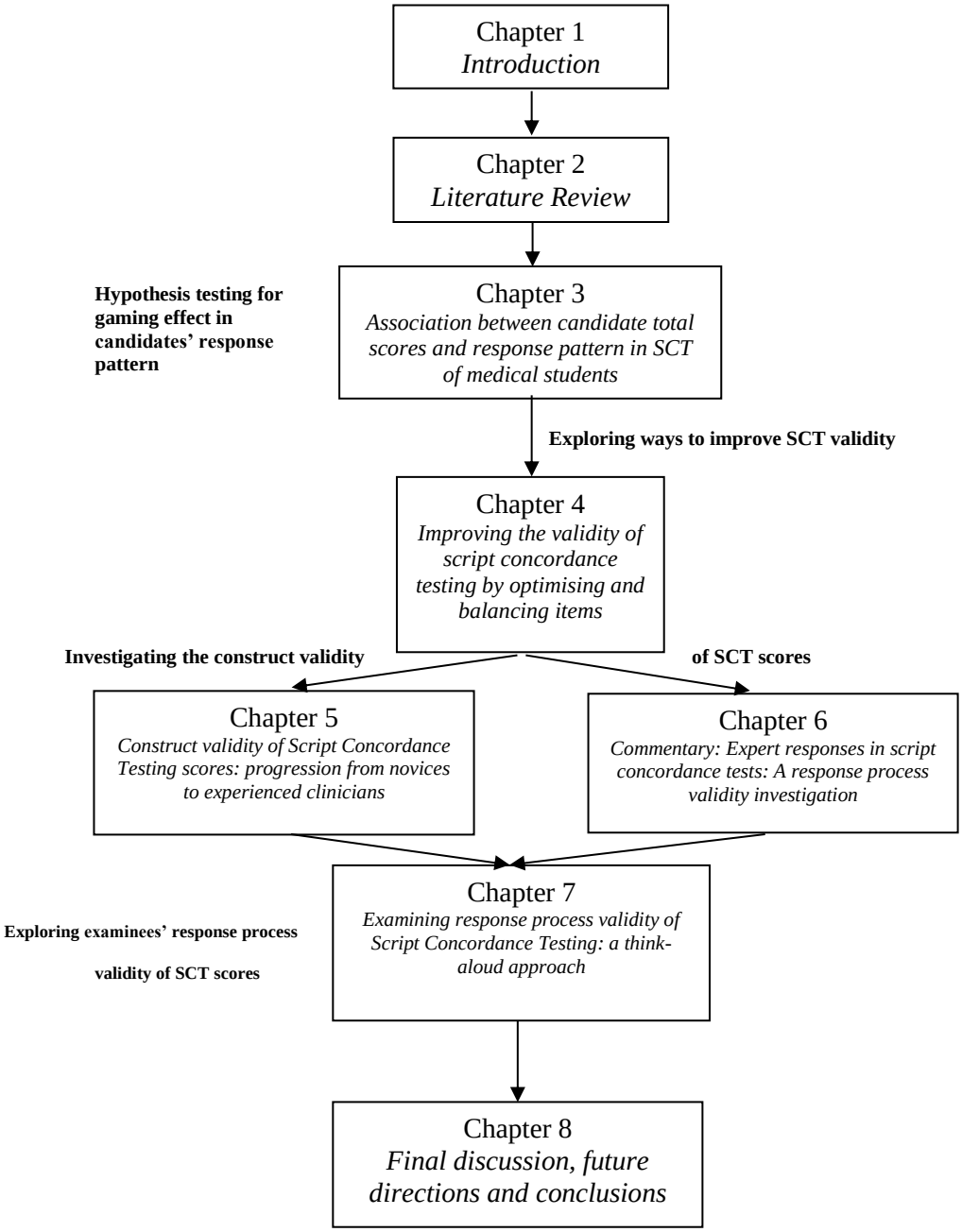
3. The fourth study and published paper (Chapter 5) investigated whether clinical reasoning skills improved with increasing clinical experience. It examined whether the SCT scores of medical students in the study improved as they progressed from Year 3 to Year 4 of the Medical Programme, and whether there was further progression of SCT scores for junior doctors and experienced practising General Practitioners (GPs). (58) Evidence for the construct validity of SCT, as measured by improvement in SCT scores as undergraduate and postgraduate trainees progress through clinical training to practice, has not previously been reported in the literature.

4. In response to a recent publication raising concerns about the validity of SCT scores due to variations in expert panel's response processes in answering SCT questions, an invited Commentary paper (Chapter 6) discussed and proposed ways to address some of the concerns. (59) Specifically, discussion on using SCT as a form of assessment 'for' learning in addition to assessment 'of' learning was presented.

5. The final study and paper in Chapter 7, examined the response process validity of SCT scores by using a written ‘think-aloud’ approach to investigate medical students’ response process in answering SCT items. It investigated the underlying thought process of the candidate’s clinical reasoning in answering the SCT questions. This ‘think-aloud’ approach enhanced the feedback provided to students after a formative SCT examination, allowing them to compare their clinical reasoning to that of the expert panel. (60)

Chapter 8 discusses the main findings from this PhD project in the context of existing literature. After consideration of the limitations, of the thesis research, future directions for ongoing research into the use of SCT and the validity of SCT scores in assessing clinical reasoning competency in medical education, are proposed.

1.7 Presentation of thesis: Flow chart of the structure and chapters



Chapter 2: Literature Review – Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine

This chapter contains the literature review titled "Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine" published in the Hong Kong Medical Journal (ERA Journal). 2015;21(5):455-61.

Statement of Contribution by Others: please refer to Appendix 1

Foreword

The following published article reviewed the literature on the historical development of SCT as a tool to assess clinical reasoning in Medicine. It highlighted the structure, scoring, standard-setting, and the importance of reliability and validity of SCT. This provided the background information on the use of SCT in medical education and identified some of the current issues related to its validity, at the commencement of the PhD.

Approval granted from Journal Editor for inclusion in the Thesis.

Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine

SH Wan *

ABSTRACT

The script concordance test is a relatively new format of written assessment that is used to assess higher-order clinical reasoning and data interpretation skills in medicine. Candidates are presented with a clinical scenario, followed by the reveal of a new piece of information. The candidates are then asked to assess whether this additional information increases or decreases the probability or likelihood of a particular diagnostic, investigative, or management decision. To score these questions, the candidate's decision in each question is compared with that of a reference panel of expert clinicians. This review focuses on the development of quality script concordance questions, using expert panellists to score the items and set the passing score standard, and the challenges in the practical implementation (including pitfalls to avoid) of the written assessment.

Hong Kong Med J 2015;21:455–61
DOI: 10.12809/hkmj154572

SH Wan *, MB, ChB, MRCP (Edin)

School of Medicine Sydney, University of Notre Dame, 160 Oxford Street, Darlinghurst, NSW 2010, Australia

* Corresponding author: michael.wan@nd.edu.au

This article was published on 28 Aug 2015 at www.hkmj.org.

Introduction

Script concordance test (SCT) is a relatively new format of written assessment to assess higher-order clinical reasoning and data interpretation skills of medical candidates.[1]

In recent years, universities and postgraduate colleges worldwide have used SCT for both formative and summative assessment of clinical reasoning in various medical disciplines including paediatric medicine, paediatric emergency medicine, neurology, orthopaedics, surgery, and radiology.[2-8] In the classic written assessment, multiple-choice questions (MCQ) and short-answer questions (SAQ) usually examine the candidates' simple knowledge recall at the lowest 'knows' level of the Miller's Pyramid (Fig 1).[9,10] Questions in SCT are able to test candidates at the higher order of thinking at the 'knows how' and even 'shows how' level. It is a unique assessment tool that targets the essential clinical reasoning and data interpretation skills in a very authentic way that reflects the element of 'uncertainty' in real-world clinical scenarios prevalent in clinical practice. This is the key aspect of clinical competency that enables medical graduates or fellows in training to link and transfer their mastery of declarative clinical knowledge and skills into clinical practice in a real clinical setting. Recent literature reports the value of using SCT to assess other areas of disciplines where classic questions are difficult to develop, for example, in assessing medical ethical principles and professionalism.[11]

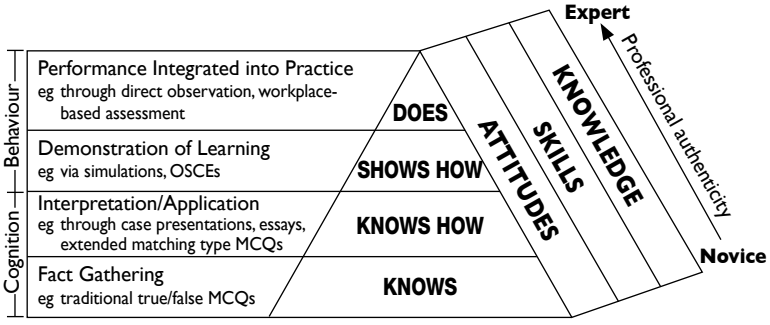


FIG 1. Miller's Pyramid[9,10]
Reprinted with permission from Dr Ramesh Mehay
Abbreviations: MCQs = multiple-choice questions; OSCEs = objective structured clinical examinations

The structure and format of script concordance test

In SCT, candidates are presented with a clinical vignette/scenario, followed by the reveal of a new piece of information. The candidates are then asked to assess whether this additional information increases or decreases the probability or likelihood of the suggested provisional diagnosis, and increases or decreases the usefulness/appropriateness of a proposed investigation or management option. The process reflects everyday real-world decision-making processes where clinicians retrieve their 'illness scripts' or network of knowledge (about similar patient problems and presentations stored in their memory) when faced with uncertainty in a clinical presentation. This enables them to determine the follow-on diagnosis and management options most appropriate to the situation. As further clinical encounters are experienced, the scripts are updated

and refined.[12] Script concordance test assesses the candidates' clinical reasoning and data interpretation ability in the context of uncertainty, particularly involving ill-defined patient problems in clinical practice.[13] Sample SCT questions in Table 1 illustrate the structure and format of the SCT questions. As the clinical scenario unfolds, additional data such as clinical photos, radiological images, or audiovisual material can also be presented to enhance the authenticity of the scenarios.[5,14,15]

In scenario A in Table 1, the 'clinical vignette' is that of a 22-year-old woman who presents to the Emergency Department with severe abdominal pain. A piece of 'new information' is then revealed that her serum beta–human chorionic gonadotropin (β-HCG) is normal. The candidate is asked whether this additional information makes the 'diagnosis' of ectopic pregnancy: much less likely (-2), less likely (-1), neither more nor less likely (0: no effect on the likelihood), more likely (+1), or much more likely (+2). The next piece of new information (independent of the first one) is that the examination shows marked guarding and rigidity of the abdomen and the candidate is asked to determine the likelihood of a diagnosis of acute appendicitis.

In scenario B in Table 1, a similar format is used to assess the appropriateness of ordering an investigation in relation to the respective piece of additional information. The first question asks for the appropriateness of ordering a computed tomographic scan of the abdomen for a 16-year-old girl who presents with acute abdominal pain if her last menstrual period was 8 weeks ago.

In scenario C in Table 1, the focus is on the usefulness of different management options after being presented with different pieces of new information related to the clinical vignette.

In preparing candidates to answer the questions, it is crucial to emphasise that each piece of new information is independent of the previous piece but in the same clinical setting. For example, in scenario A, when answering the second question given that she has guarding and rigidity in the abdomen, she does NOT have a serum β-HCG test done.

With respect to the likelihood descriptors used in the SCT questions for the diagnosis type of scenario, the preference is to use the option of "much less likely (-2)" rather than "ruling out the diagnosis"; and "much more likely (+2)" rather than "almost certain/definite diagnosis". This will allow candidates to use the full range of the five options. In the practice of medicine, there are usually few situations wherein a diagnosis can be confidently excluded or definitely diagnosed with a few pieces of information provided.[3]

There are nonetheless limitations to the design and format of SCT. Candidates cannot seek additional information to that given in the question; the scenario is only a snapshot of the clinical encounter without the comprehensive history, physical examination, and investigations that would be particularly desirable in an ambiguous clinical situation.[16]

Scoring script concordance test

To score these questions, the candidate's decision in each question is compared with that of a reference panel of expert clinicians. Each member of the panel attempts the same set of questions and the answers are recorded. As there is no single best correct answer to the question, a full (1) mark will be awarded if the candidate's decision concurs (hence the name 'concordance') with the majority of the expert panel. A proportional (partially credited or weighted) score (<1) will be given if the candidate's decision concurs with the minority of the panel. The candidate will score a '0' if no panellist chooses this option.[3] The formula to calculate the weighted scores is shown in Table 2.

There are other scoring methods reported in the literature where a consensus-based single-answer scoring method or 3-point Likert scale scoring method is employed to determine the candidate scores.[4,17]

Selecting the reference panel

In general, a panel of 10 to 15 expert members relevant to the discipline is recommended to produce credible and reliable scores.[18] The inter- and intra-rater reliability in the SCT panel have been shown to be good.[19]

The composition of the panel should include clinical teachers and academics who are familiar with the curriculum and experts in the field relevant to the discipline tested. Studies have shown that using general practitioners (GPs) in the panel may produce similar mean scores to specialists but with a wider standard deviation.[3]

TABLE 1. Sample questions of script concordance test

Clinical scenario								
A: A 22-year-old woman presents to the Emergency Department with severe abdominal pain.								
	If you were thinking of...	**and then you find that…**	**this hypothesis becomes …**					
1	Ruptured ectopic pregnancy	Her serum β-HCG is negative	**A** -2	**B** -1	**C** 0	**D** +1	**E** +2	-2: Much less likely -1: Less likely 0: Neither more nor less likely +1: More likely +2: Much more likely
2	Acute appendicitis	On abdominal examination, there is marked guarding and rigidity	**A** -2	**B** -1	**C** 0	**D** +1	**E** +2	
3	Acute cholecystitis	Her temperature is 36.8°C	**A** -2	**B** -1	**C** 0	**D** +1	**E** +2	
B: A 16-year-old girl is brought to the Emergency Department by her parents. She has been vomiting and complains of generalised abdominal pain.								
	If you were thinking of ordering the following…	**and then you find that…**	**then your plan of action becomes …**					
4	CT abdomen	Her last menstrual period was 8 weeks ago	**A** -2	**B** -1	**C** 0	**D** +1	**E** +2	-2: Much less appropriate -1: Less appropriate 0: Neither more nor less appropriate +1: More appropriate +2: Much more appropriate
5	Laparoscopy	CT abdomen is normal	**A** -2	**B** -1	**C** 0	**D** +1	**E** +2	
6	CT abdomen	Her blood glucose level is 32 mmol/L (reference range, 3.5-7.0 mmol/L)	**A** -2	**B** -1	**C** 0	**D** +1	**E** +2	
C: A 55-year-old woman with previous asthma presents with acute shortness of breath. She is afebrile. You find she has a diffuse expiratory wheeze.								
	If you were thinking of …	**and then you find that…**	**then your plan of action becomes …**					
7	Giving morphine for her distress	Her PO_2 is 55 mm Hg and her PCO_2 is 60 mm Hg	**A** -2	**B** -1	**C** 0	**D** +1	**E** +2	-2: Much less appropriate -1: Less appropriate 0: Neither more nor less appropriate +1: More appropriate +2: Much more appropriate
8	Giving hydrocortisone intravenously	Her blood glucose is 24.2 mmol/L	**A** -2	**B** -1	**C** 0	**D** +1	**E** +2	
9	Giving 5 mg salbutamol by nebuliser	Her pulse rate is 120 bpm	**A** -2	**B** -1	**C** 0	**D** +1	**E** +2	

Abbreviations: bpm = beats per minute; β-HCG = beta–human chorionic gonadotropin; CT = computed tomography; PCO_2 = partial pressure of carbon dioxide; PO_2 = partial pressure of oxygen

TABLE 2. The formula to calculate the weighted scores

Score key	-2	-1	0	+1	+2
No. of panellists choosing the answer (out of 10)	7	2	1	0	0
Formula	7/7	2/7	1/7	0/7	0/7
Candidate score	1	0.29	0.14	0	0

A recent study, however, raised concerns about the reference standard and judgement of the expert panel. The study compared 15 emergency medicine consultants' judgement scores with evidence-based likelihood ratios. The results showed that 73.3% of the mean judgement was significantly different to the corresponding likelihood ratios, with 30% overestimation, 30% underestimation, and 13.3% with diagnostic values in the opposite direction.[20] Other studies raised concerns about the possibility of outdated clinical knowledge and cognitive bias in the experts' decision-making.[21,22] Evidence of context specificity has also been highlighted whereby the agreement between SCT scores derived using different scoring keys with expert reference panels from a different context (hospitals and specialty) was poor.[23]

Implementation of script concordance test in formative and summative assessments

The structure and layout of the SCT questions can easily be implemented in the usual pen and paper-based or online electronic format. Candidates answer each question (with five options) using a standardised answer sheet to facilitate computer scanning and scoring or directly online using the computer.

It is often difficult to get busy clinicians to meet together face-to-face to answer the questions. By uploading the questions online, the panellists can attempt them anytime and make the questions available through a secure online platform. The response data can then be collated and the weighted

scores for responses on each score scale calculated.[3]

After capturing the candidates' responses for all items, scoring of responses for each question can then be performed using the formula described above. This will ensure a rapid turnaround time that will be very effective in the assessment process.

For formative assessment purposes, expert panel consensus scores are provided to the candidates, followed by expert clinicians explaining and discussing the options in each scenario with the candidates for constructive feedback. Script concordance test can also be used to identify borderline students with suboptimal clinical reasoning skills for appropriate remedial measures such as bedside teaching, tutorials, or clinical simulations.[24]

For summative assessment purposes, particularly where there is not a large pool of SCT items, it is important to avoid constructing irrelevant variance in SCT scores, by not releasing or discussing post-examination, the expected responses (based on expert panel's responses), and the associated score for each of the answer options in SCT items.

Unlike MCQ where there is only one single best answer that candidates could memorise and disseminate after the examination, the partial credit scoring model applied in SCT, where multiple answer options are accepted and each carries a fraction or all of the allocated mark, has to a certain extent rendered sharing of 'correct' answers after the examination difficult.

Developing quality script concordance test questions

Each clinical scenario has to be authentic and the presentation represents a realistic clinical encounter that is relevant to the specific discipline, preferably with a certain degree of uncertainty. The new information presented needs to stimulate the candidate to re-consider and re-evaluate how that particular piece of new information will affect the likelihood of the initial diagnosis, or appropriateness of initial planned investigation or management option. This will ensure the content validity in the SCT questions.

Particular care should be taken to develop options that will attract the full range of the five options available for the candidate to choose from. In other words, the additional pieces of new information should result in the consideration of -2 and +2 as well as -1, 0, and +1 options. A test-wise candidate might choose to consider only the options of -1, 0 and +1 if they notice that most panel consensus answers with a full score of 1 mark usually fall within these three options rather than also covering the -2 and +2.[25] As a result, developing good-quality SCT questions is not easy. Care should be taken to develop clinical scenarios that do not focus solely on factual recall but involve a reasoning process with elements of uncertainty that will likely attract responses that spread across the 5-point Likert scale.[26]

Reliability and validity of script concordance test as an assessment tool

The reliability of SCT as an assessment tool has been investigated.[2,6] A 60- to 90-minute examination will produce a Cronbach's alpha of 0.70 to 0.85.[7,25,27,28] There are concerns, however, about inter-panellist errors in SCT; the use of Cronbach's alpha in measuring the reliability of the test where partial credit model of scoring is used, ie multiple options/responses are awarded either a full or fraction of allocated mark; and case scenarios that could create inconsistencies among items.

As an assessment tool, SCT has been shown to be valid in assessing clinical reasoning.[13,14,19,28] Studies have shown that SCT scores correlate well with other assessment scores from the clinical years of the candidates.[2]

The construct validity of SCT questions can be examined by correlating the scores with the level of training to predict future performance on clinical reasoning. A recent study has compared the progression of clinical reasoning skills of medical students with those of a group of practising GPs who are also their Problem Based Learning group tutors.[29] Another study showed that there was a statistically significant gain in SCT performance over a 2-year period in two different cohorts of medical students using the same set of 75-item SCT.[26] There was significant progression of clinical reasoning skills from medical students at the novice level through to practising GP clinicians, reflected by the higher scores in the GP group attempting the SCT questions. Empirical evidence supporting the construct validity based on progression of SCT scores with clinical experience from undergraduate students to postgraduate training has also been reported.[2,5,24,30,31] The construct validity of SCT has been questioned because of the logical inconsistencies as a result of partial credit scoring methodology making it possible for a hypothesis to be simultaneously more likely and less likely.[32] Nonetheless, a certain degree of variability in panel scores has been shown to be a key determinant of the discriminatory power of the test and allows richness of thinking about clinical cases.[33,34] Another study found that 27% of residents in one SCT administration scored above the expert panel's mean, which may indicate issues with the construct validity, particularly in the credibility and validity of the scoring key and hence the resulting SCT scores interpretation.[33]

Test-wise candidates would select the answers to be around '0' rather than '-2' or '+2' if they noticed

that most panellist scores did not fall in the 'extreme' (-2 or +2) range due to the construct of the SCT questions and options. This could be avoided by first using the option descriptor of "much less likely (-2)" and "much more likely (+2)" rather than "ruling out the diagnosis" and "almost certain/definite diagnosis" as described in the format of SCT section above.[19] Second, when collating the SCT questions into an examination paper, one could select a relatively equal number of items with both 'extreme' answers as well as median answers. Recent data have shown that by employing the above strategies in developing the paper, candidates who chose '0' for all the questions would score only around 25% in the SCT examination and would gain no advantage (unpublished data). This is in contrast to the finding of another study wherein candidates who chose the midpoint scale ('0') performed better than the average candidate.[32]

The correlation of SCT scores with other modalities of assessment would be expected to be low as SCT is designed to measure clinical reasoning rather than factual or knowledge recall. The correlation coefficient between SCT and MCQ was poor (r=0.22), and that between SCT and extended matching questions (EMQ) was 0.46.[4]

Collating and moderating the expert reference panellist responses

In collating the SCT questions for use in a summative examination, appropriate clinical scenarios/vignettes with the related diagnoses, investigations, and management should be selected according to the blueprint of the assessment. The clinical topics should have a good spread and represent core areas of learning that are relevant to the curriculum and appropriate to the level of training of the candidates.

In reviewing the expert panel responses to each question, bi-modal and uniform divergence responses should trigger a detailed scrutiny of the clinical vignette and the options. In the case of bi-modal response (Fig 2a), the panel has an equal split of the best option between -2 and +2. This usually results from an error in the question or a controversial investigation or management option with discordant 'expert opinions'. A modification of the question and re-scoring will usually solve this issue. If re-scoring results in the same bi-modal response, the question should be discarded for scoring in the examination. In the case of uniform divergence responses (Fig 2b), there is an equal spread in the number of members choosing all the five options. This usually signifies a non-discriminating question and the item should again be discarded. A discrete outlier response (Fig 2c) usually represents an error in the particular panellist's decision or 'clicking the wrong answer accidentally' when the member should have answered -2 instead of +2. The ideal pattern would be relatively close convergence with some variation (Fig 2d).[3]

As mentioned previously, the set of questions in the SCT examination should be selected in such a way that there are similar numbers of full marks in each option across the five options. This will avoid the test-wise candidates being advantaged by selecting only the -1, 0, or +1 options and avoiding the extreme options of -2 and +2.[3] By employing this strategy to select questions that cover the full 5-point Likert scale, test-wise students will only score 25% in the SCT examination if they choose the response of '0' for all questions (unpublished data) compared with 57.6% in another cohort sitting a SCT test without the specific question selection process.[32]

Standard setting the pass/fail cutoff score

In setting the pass/fail cutoff score of the SCT questions, the expert panels' mean scores and standard deviations are chosen to guide the process. This is calculated by asking all the members of the panel to attempt the same set of SCT questions and their responses are then scored accordingly. The borderline score of the undergraduate students is usually set at 3 to 4 standard deviations below the expert panel's mean score.[3,35] Studies have shown that using recent graduates or fellows in training might result in a mean score that is closer to the students' mean and therefore a smaller number of standard deviations would be more appropriate.[3]

Other methods of standard setting include using the single correct answer method.[29,36] Standard setting of a pass/fail cutoff score is an area that warrants ongoing research to inform and improve the practice of using SCT as a summative assessment tool for clinical data interpretation and decision-making skills.

The use of script concordance test in the Asia-Pacific region and its limitations

Examinations using SCT have been successfully implemented in the school-entry medical schools in Indonesia, Singapore, Taiwan, and Australia[3,7,36,37]; and in graduate-entry medical schools in Australia.[29,38] Such test has the potential to supplement MCQ and SAQ to test the higher-order thinking of medical candidates to allow a more robust overall written assessment in the assessment programme. In fact, SCT is one of the few currently available assessment tools for clinical reasoning in a written format.[28] It can be implemented relatively easily in the paper-based format or online. Initial

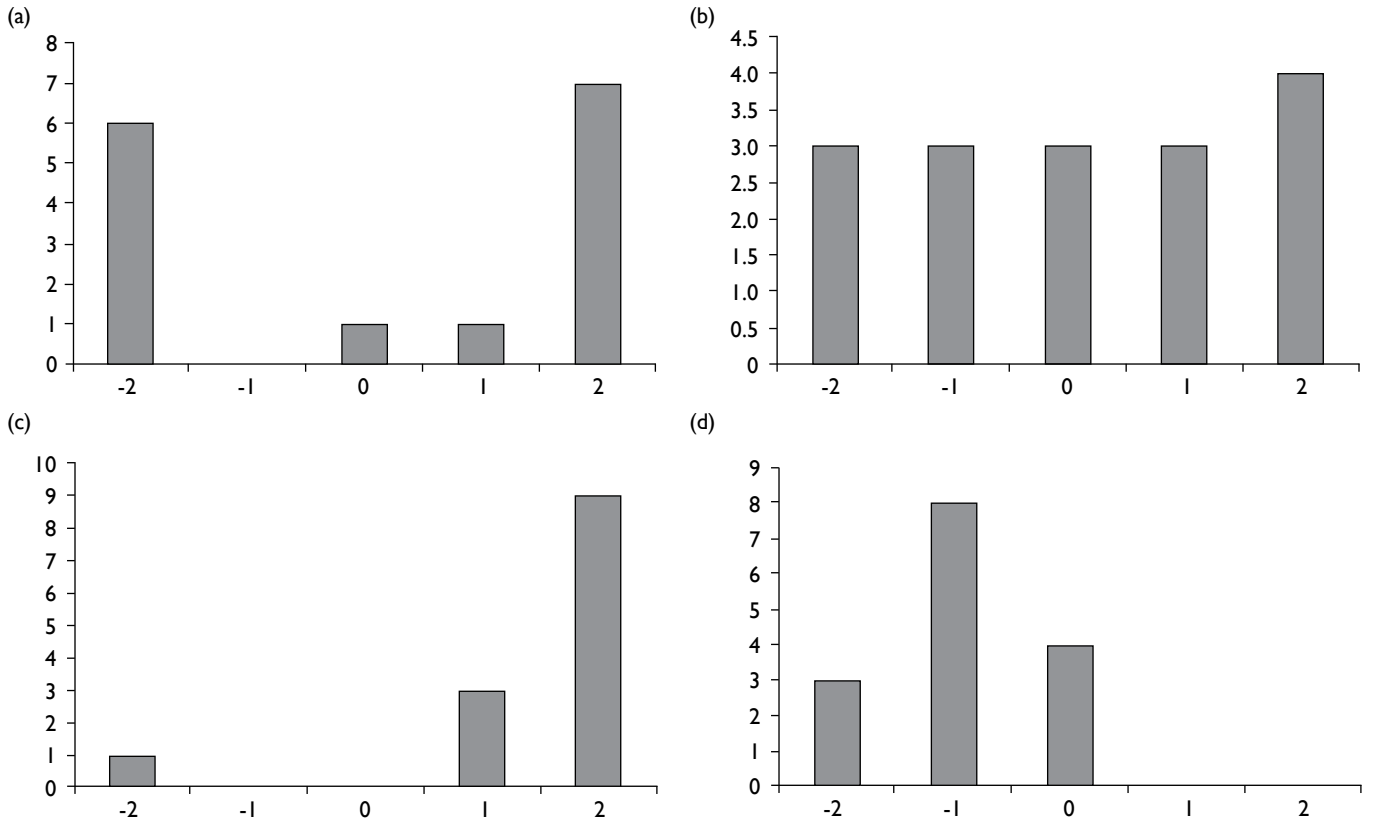


FIG 2. Number of responses from the expert panel to script concordance test questions
(a) Bi-modal response, (b) uniform divergence response, (c) discrete outlier response, and (d) ideal response

pilot examinations can be set as a formative exercise to enhance candidates' feedback and learning.[24] Further collaboration with other institutions to develop, score, and share question items can ensure effective and efficient delivery of such examinations.

Limitations to the widespread usage of SCT could be due to: difficulties in developing good-quality SCT clinical scenarios, concerns about the validity of the test, recruiting a sufficient number of appropriate expert clinicians for the reference panel, lack of a general consensus in setting the borderline pass mark, and the candidates' familiarity with the question format.[3,24,25,28,32,34]

Conclusions

This article attempts to review the current use of SCT in assessing clinical reasoning and data interpretation skills in undergraduate and postgraduate medicine. The empirical evidence reported for the reliability and validity of SCT scores from existing literature seems encouraging. Approaches to develop quality items, moderation of expert panel scoring and these post-hoc quality assurance measures, and optimisation of scoring scale will to a certain extent mitigate the threat to the validity of SCT score interpretation and its use for summative examination purposes. Combining SCT (testing the clinical reasoning and data interpretation skills with authentic written simulations of ill-defined clinical problems set at the 'knows how' level) with MCQ/SAQ/EMQ (testing the 'knows' and 'knows how'), objective structured clinical examination (testing 'shows how'), and workplace-based assessment (testing the 'does') in the medical curriculum will enhance the robustness and the credibility of the assessment programme.

Further research into the use of SCT in both undergraduate and postgraduate medical education is warranted, particularly on standard setting for the pass/fail cutoff score and best practices that may help reduce the threat to the validity of SCT scores.

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. Teach Learn Med 2000;12:189-95.
2. Carrière B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a Script Concordance Test. Ann Emerg Med 2009;53:647-52.
3. Duggan P, Charlin B. Summative assessment of 5th year medical students' clinical reasoning by Script Concordance Test: requirements and challenges. BMC Med Educ 2012;12:29.

4. Kelly W, Durning S, Denton G. Comparing a script concordance examination to a multiple-choice examination on a core internal medicine clerkship. Teach Learn Med 2012;24:187-93.
5. Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. The Script Concordance Test: a new tool assessing clinical judgement in neurology. Can J Neurol Sci 2009;36:326-31.
6. Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? Am J Surg 2007;193:248-51.
7. Soon D, Tan N, Heng D, Chiu L, Madhevan M. Neurologists vs emergency physicians: reliability of a neurological script concordance test in a multi-centre, cross-disciplinary setting. Neurology 2014;82(10 Suppl):327.
8. Talvard M, Olives JP, Mas E. Assessment of medical students using a script concordance test at the end of their internship in pediatric gastroenterology [in French]. Arch Pediatr 2014;21:372-6.
9. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65(9 Suppl):S63-7.
10. Mehay R. The essential handbook for GP training and education. Chapter 29. Assessment and competence. Available from: http://www.essentialgptrainingbook.com/chapter-29.php. Accessed 10 May 2015.
11. Foucault A, Dubé S, Fernandez N, Gagnon R, Charlin B. Learning medical professionalism with the online concordance-of-judgment learning tool (CJLT): A pilot study. Med Teach 2014 Oct 22:1-6. Epub ahead of print.
12. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implications. Acad Med 1990;65:611-21.
13. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance testing: a review of published validity evidence. Med Educ 2011;45:329-38.
14. Brazeau-Lamontagne L, Charlin B, Gagnon R, Samson L, Van Der Vleuten C. Measurement of perception and interpretation skills during radiology training: utility of the script concordance approach. Med Teach 2004;26:326-32.
15. Collard A, Gelaes S, Vanbelle S, et al. Reasoning versus knowledge retention and ascertainment throughout a problem-based learning curriculum. Med Educ 2009;43:854-65.
16. Lineberry M, Kreiter CD, Bordage G. Script concordance tests: strong inferences about examinees require stronger evidence. Med Educ 2014;48:452-3.
17. Williams RG, Klamen DL, White CB, et al. Tracking development of clinical reasoning ability across five medical schools using a progress test. Acad Med 2011;9:1148-54.
18. 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? Med Educ 2005;39:284-91.
19. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: a systematic review. Med Educ 2012;46:552-63.
20. Ahmadi SF, Khoshkish S, Soltani-Arabshahi K, et al. Challenging script concordance test reference standard by evidence: do judgments by emergency medicine consultants agree with likelihood ratios? Int J Emerg Med 2014;7:34.
21. Ramos K, Linscheid R, Schafer S. Real-time information-seeking behavior of residency physicians. Fam Med 2003;35:257-60.
22. Norman GR, Eva KW. Diagnostic error and clinical reasoning. Med Educ 2010;44:94-100.
23. Tan N, Tan K, Ponnamperuma G. Expert clinical reasoning is not just local but hyperlocal—insights into context specificity from a multicentre neurology script concordance test. Neurology 2015;84(14P4):191.
24. Ducos G, Lejus C, Sztark F, et al. The Script Concordance Test in anesthesiology: Validation of a new tool for assessing clinical reasoning. Anaesth Crit Care Pain Med 2015;34:11-5.
25. See KC, Tan KL, Lim TK. The script concordance test for clinical reasoning: re-examining its utility and potential weakness. Med Educ 2014;48:1069-77.
26. Humbert AJ, Miech EJ. Measuring gains in the clinical reasoning of medical students: longitudinal results from a school-wide script concordance test. Acad Med 2014;89:1046-50.
27. Gagnon R, Charlin B, Lambert C, Carrière B, Van der Vleuten C. Script concordance testing: more cases or more questions? Adv Health Sci Educ Theory Pract 2009;14:367-75.
28. Nouh T, Boutros M, Gagnon R, et al. The script concordance test as a measure of clinical reasoning: a national validation study. Am J Surg 2012;203:530-4.
29. Wan SH. Using Script Concordance Testing (SCT) to assess clinical reasoning—the progression from novice to practising general practitioner. Proceedings of the 11th Asia Pacific Medical Education Conference; 2014 Jan 15-19; Singapore.
30. Charlin B, van der Vleuten C. Standardized assessment in contexts of uncertainty: The script concordance approach. Eval Health Prof 2004;27:304-19.
31. Lambert C, Gagnon R, Nguyen D, Charlin B. The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. Radiat Oncol 2009;4:7.
32. Lineberry M, Kreiter C, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ 2013;47:1175-83.
33. Charlin B, Gagnon R, Lubarsky S, et al. Assessment in the context of uncertainty using the script concordance test: more meaning for scores. Teach Learn Med 2010;22:180-6.
34. Lubarsky S, Gagnon R, Charlin B. Scoring the Script Concordance Test: not a black and white issue. Med Educ 2013;47:1159-61.
35. Wan SH, Clarke R. Using a clinician panel to set the borderline mark for Script Concordance Testing (SCT) to assess clinical reasoning for graduating medical candidates. Proceedings of the 8th International Medical Education Conference; 2013 March 13-15; Kuala Lumpur, Malaysia.
36. Irfannuddin I. Knowledge and critical thinking skills increase clinical reasoning ability in urogenital disorders: a Universitas Sriwijaya Medical Faculty experience. Med J Indonesia 2009;18:53-9.
37. Tsai TC, Chen DF, Lei SM. The ethics script concordance test in assessing ethical reasoning. Med Educ 2012;46:527.
38. Ingham AI. The great wall of medical school: a comparison of barrier examinations across Australian medical schools. Aus Med Student J 2011;2:5-8.

Synopsis of Chapter 2

Chapter 2 covers a general overview of published literature on the development and implementation of Script concordance testing (SCT) in medical education. A brief discussion on the importance of moderation of expert panel responses and item selection to improve the quality of SCT was presented. The use of SCT in the Asia-Pacific region was also presented. The limitations and threats to the validity of SCT were highlighted. Building on the above background, the next chapter will report findings from an investigation on the real word actual characteristics of medical students' response patterns in the summative high stakes examinations, to confirm or refute the alleged threats to the validity of the SCT scores where students avoid extreme response options as an attempt to gain a higher score.

Chapter 3: Association between candidate total scores and response pattern in script concordance testing of medical students

This chapter contains the paper titled "Association between candidate total scores and response pattern in script concordance testing of medical students" published in Focus on Health Professional Education: A Multi-disciplinary Journal (ERA Journal). 2017;18(2):26-35.

Statement of Contribution by Others: please refer to Appendix 2

Foreword

The following published article investigated the response patterns of 6 cohorts of clinical year medical students in the study context, to support or refute the reported concerns in the literature that students would avoid extreme response options as a test wise strategy in the SCT examination.

This is the peer reviewed version of the above article, which has been published in final form at Focus on Health Professional Education: A Multi-disciplinary Journal (ERA Journal). 2017;18(2):26-35 (http://dx.doi.org/10.11157/fohpe.v18i2.145)

Association between candidate total scores and response pattern in script concordance testing of medical students

S. H. Wan[1], P. Duggan[2], E. Tor[1] & J. N. Hudson[2]

Abstract

Introduction: The script concordance test (SCT) aims to test clinical decision making and clinical reasoning. This study is a preliminary attempt to understand an alleged test-taking strategy where students avoid extreme response options, potentially threatening the validity of SCT scores. We investigated whether there is a significant association between the propensity to avoid the extreme response options and candidates' overall SCT scores.

Methods: The SCT scores of 660 clinical-year medical students (six cohorts from 2013–2015) were analysed for a possible association with candidates' response pattern. The proportion of middle range response options was calculated. Propensity to avoid extreme response options is defined as a response pattern with 15% or more of middle-range responses compared to those of the expert reference panel. The distribution for candidates with propensity to avoid the extreme options was further investigated using chi-square statistics for possible association with their overall SCT results.

Results: Fifty-five percent of the students from the lowest quartile, compared to 30% from the top quartile, had shown a propensity to avoid the extreme options. The differences were statistically significant ($p < 0.001$) and were consistent among all six cohorts included in this study.

Conclusions: Students whose SCT scores are in the lowest quartile are more likely to avoid the extreme response options in answering SCT questions. For quality assurance in high stakes summative SCTs, it may be worthwhile to select items with expert reference panel's modal answers covering the full 5-point response options.

Keywords: medical education; script concordance; clinical reasoning; assessment.

1 School of Medicine, University of Notre Dame, Australia
2 School of Medicine, University of Adelaide, Australia

Correspondence
Siu Hong Wan
Head of Basic and Clinical Science Domain
Associate Professor, Assessment
School of Medicine
University of Notre Dame
Australia
Tel: +61 2 8204 4479
Email: michael.wan@nd.edu.au

ISSN 1442-1100

Introduction

The script concordance test (SCT) was introduced in 2000 by Charlin, aiming to assess the higher-order clinical reasoning skills of medical students (Charlin, Roy, Brailovsky, Goulet, & van der Vleuten, 2000). It is a useful assessment tool to test clinical reasoning and data interpretation skills, and has been shown to be valid (Lubarsky, Vleuten, Charlin, Chalk, & Cook, 2011).

The SCT is a written format currently in widespread use internationally to test clinical reasoning in health professional education. In recent years, the SCT has been used in various medical disciplines, such as internal medicine, paediatrics, emergency medicine, neurology, surgery, anaesthesia and radiology (Boulouffe, Doucet, Muschart, Charlin, & Vanpee, 2014; Brazeau-Lamontagne, Charlin, Gagnon, Samson, & van der Vleuten, 2004; Carrière, 2009; Drolet, 2015; Nouh et al., 2012; Tan, Tan, Kandiah, Samarasekera, & Ponnamperuma, 2011). The SCT has also been used to assess other discipline areas where classical written multiple-choice questions (MCQs) or short-answer questions (SAQs) are difficult to develop, for example, in assessing medical ethical principles and professionalism (Foucault, Dubé, Fernandez, Gagnon, & Charlin, 2015; Tsai, Chen, & Lei, 2012). While more traditional assessment formats such as MCQs and SAQs tend to assess students' lower taxonomic orders of thinking, SCT questions can be used to assess a higher order of thinking (Palmer, Duggan, Devitt, & Russell, 2010). Some forms of modified essay questions (MEQs) have been shown to fail to assess higher cognitive skills and have been replaced with a SCT examination (Duggan & Charlin, 2012; Palmer et al., 2010).

The SCT has been shown to be both valid and reliable in several studies, including a country-wide validation study (Dory, Gagnon, Vanpee, & Charlin, 2012; Lubarsky et al., 2011; Nouh et al., 2012; Wan, 2015). The reliability of a 60 to 90-minute examination had a Cronbach alpha of 0.7–0.85 (Nouh et al., 2012; See, Tan, & Lim, 2014). Evidence supporting the construct validity based on the progression of SCT performance related to the clinical experience from undergraduate students to post-graduate fellowship training has also been reported (Ducos et al., 2015; Lambert, Gagnon, Nguyen, & Charlin, 2009; Wan, 2014).

The SCT assessment format has been successfully implemented in undergraduate and graduate-entry medical schools, residency and fellowship training worldwide as well as in nursing schools (Chang et al., 2014; Dawson, Comer, Kossick, & Neubrander, 2014; Duggan & Charlin, 2012; Irfannuddin, 2009; Kow, Walters, Karram, Sarsotti, & Jelovsek, 2014; Nouh et al., 2012; Palmer et al., 2010). In fact, SCT is one of the few currently available assessment tools for clinical reasoning in the written format (Nouh et al., 2012). It can be implemented relatively easily in the paper-based format or online, and the scoring can be done electronically.

In a typical SCT question, candidates are presented with a clinical scenario followed by an additional piece of information. They are then asked for the probability of the suggested diagnosis or the appropriateness of a proposed investigation or management.

The descriptors for the response options range from ruling out/contraindicated (-2), less likely/less appropriate (-1), neither less nor more likely/appropriate (0), more likely/appropriate (+1) to definitive diagnosis/absolutely necessary (+2).

This process reflects how practising clinicians retrieve their "illness scripts" or network of previous clinical experience (about similar patient encounters) when faced with uncertainty with diagnosis, investigation or management (Lubarsky et al., 2011; Wan, 2015).

In order to allow the students to choose from the full range of the five response options, "much less likely (-2)" rather than "ruling out the diagnosis" and "much more likely (+2)" rather than "definitive diagnosis" are used in the questions in our school (Wan, 2015). Two sample SCT questions on diagnosis and management are shown in Figure 1.

To score these SCT questions, the student's decision is compared to that of a reference expert clinician panel. Students are able to score marks according to the "concordance" in the decision with the majority of the panel. A partial score is given if the decision concurs with a minority of the panel.

Clinical Scenario A
A 42-year-old women presents to the general practice with a lump in the neck which moves upward on swallowing.

	If you were thinking of …	**and then you find that …**	**this hypothesis becomes …**	
1	Multinodular goitre	The lump is smooth and measures around 3 cm in diameter	**A B C D E** -2 - 1 0 +1 +2	**-2:** much less likely **-1:** less likely **0:** neither more nor less likely **+1:** more likely **+2:** much more likely
2	Follicular carcinoma of the thyroid	A hard lymph node is palpable in the left cervical chain	**A B C D E** -2 - 1 0 +1 +2	
3	Toxic nodular goiter	His pulse rate is 60 bpm and he has no significant weigh loss	**A B C D E** -2 - 1 0 +1 +2	

Clinical Scenario B
A 45-year-old woman with a history of asthma presents with acute shortness of breath. She is afebrile. On examination, there is a diffuse expiratory wheeze.

	If you were thinking of …	**and then you find that …**	**then your plan of action becomes …**	
4	Giving morphine for her distress	Her PO2 is 55 mmHg and her PCO2 is 60 mmHg	**A B C D E** -2 - 1 0 +1 +2	**-2:** much less likely **-1:** less likely **0:** neither more nor less likely **+1:** more likely **+2:** much more likely
5	Giving hydrocortisone intravenously	Her blood glucose is 24.2 mmol/L	**A B C D E** -2 - 1 0 +1 +2	
6	Giving 5 mg salbutamol by nebuliser	Her pulse rate is 130 bpm	**A B C D E** -2 - 1 0 +1 +2	

Figure 1. Sample SCT questions.

ISSN 1442-1100

An example of using a formula to calculate the weighted scores is shown in Table 1.

Table 1
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	3	0	0	0
Formula	7/7	3/7	0/7	0/7	0/7
Student's score	1	0.43	0	0	0

Recent literature on the SCT highlighted the observation that the SCT format of aggregate partial credit scoring can be subjected to the validity threat of candidates' test-taking strategy of simply avoiding the extreme response options (Lineberry, Kreiter, & Bordage, 2013). This is similar to the response style coaching strategies described in situational judgment tests that could increase the candidates' scores significantly (Cullen, Sackett, & Lievens, 2006; McDaniel, Psotka, Legree, Yost, & Weekley, 2011). Candidates might choose to avoid the extreme response options (-2 or +2) thinking that the probability of these responses being correct would be low, or they might have a lack of confidence in choosing such extreme options.

Aims

In the present study, we investigated whether or not there is a significant association between the propensity to avoid the extreme response options in SCT (-2 or +2) and the overall SCT scores.

Methods

Participants

In 2013–2015, SCT examinations were implemented in our graduate-entry medical school in NSW, Australia. We collected de-identified data from six clinical SCT written examinations undertaken by three successive cohorts of penultimate-year clinical students and three successive cohorts of final-year clinical students (n = 660). A set of 40 SCT items was given in each examination. The reference panels consisted of clinician experts who were actively involved in teaching the students, general practitioners and academics. Scoring of the items was done according to the formula described in Table 1.

Analysis

We have operationalised propensity in "avoiding the extreme response options" as cases where a candidate's proportion of answers in the middle range (-1, 0, +1) for all 40 items in the SCT was 15% higher than that of the reference panel's. For example, if the reference panel's response pattern showed 50% of responses in the middle range (-1, 0, +1) in a SCT, then if a student's response pattern showed 67.5% of the answers chosen were in the middle range (-1, 0, +1), the student would be deemed to be adopting a test-taking strategy in avoiding the extreme options (-2 or +2).

De-identified data in the form of candidates' response pattern in individual SCT items, their total SCT scores, as well as the response data from the expert reference panel were collated and analysed. The proportion of responses to SCT items in the middle-range response options (i.e., -1, 0 and +1) for individual candidates were calculated. They were compared with an expert reference panel's responses, to identify cases of avoidance of extreme-response options.

Chi-square test of association between propensity in avoiding extreme options by candidates and their actual performance in SCT, i.e., the quartile where their overall SCT scores were located within the cohort, was analysed using IBM SPSS© package version 23.

Ethics approval was given by the Human Research Ethics Committee of the University of Notre Dame, Australia.

Results

A total of 660 clinical-year students from six cohorts in the school (three from final year and three from the third year in the four-year medical course) sat the SCT examination.

Using a chi-square test of independence to compare the frequency of avoidance of extreme-response options and the quartile of candidates' overall performance in SCT, a significant association was found (χ^2 (3, 660) = 26.29, $p < 0.001$) (Table 2). Candidates whose SCT scores were in the lowest (first) quartile were more likely to avoid the extreme response options (55%) than other students. This was followed by students in the second quartile (45%) and then students in the third quartile (33%). Students whose SCT scores were in the top quartile had the lowest incidence of avoidance of extreme-response options (30%).

Table 2
Chi-square Test of Independence Between Candidates' Avoidance of Extreme Responses and Percentile Rank of Their Overall SCT Performance

	Percentile Rank of SCT Scores (pooled data from 2013–2015 cohorts)					**Chi-square test of association**	
Avoidance of Extreme Response Options	25th percentile rank and below *(i.e., lowest 25% of SCT scores in cohort)* *Count (%)*	25th percentile rank to 50th percentile rank *Count (%)*	50th percentile rank to 75th percentile rank *Count (%)*	75th percentile rank and above *(i.e., highest 25% of SCT scores in cohort)* *Count (%)*	Total N	X^2 (df)	*p*
	(n = 165)	(n = 165)	(n = 165)	(n = 165)			
Yes	90 (54.55)	74 (44.85)	55 (33.33)	49 (29.70)	**660**	26.29 (3)	< 0.001
No	75 (45.45)	91 (55.15)	110 (66.67)	116 (70.30)			

The aforementioned observation from the chi-square analysis of pooled data from 2013–2015 was also evident in the data within each of the cohorts (2013 to 2015). This is reported in Table 3 and Figure 2.

Table 3
Chi-square Test of Independence Between Candidates' Avoidance of Extreme Responses and Percentile Rank of Their Overall SCT Performance—by Cohort (2013 to 2015)

		Avoidance of extreme Response Options		Chi-square test of association	
		Yes	No	Total *(N)*	χ^2 *(df,N)*
Quartile/ Percentile Rank of SCT Scores— 2013 cohort	*Lowest quartile for SCT scores (i.e., lowest 25% of SCT scores) Count (%)*	14 (26.42)	39 (73.58)	212	χ^2 (3,212) = 8.58 p = 0.035
	2nd quartile for SCT scores (25th percentile rank to 50th percentile rank) Count (%)	6 (11.32)	47 (88.68)		
	3rd quartile for SCT scores (50th percentile rank to 75th percentile rank) Count (%)	4 (7.55)	49 (92.45)		
	Top quartile for SCT scores (i.e., highest 25% of SCT scores) Count (%)	7 (13.21)	46 (86.79)		
Quartile/ Percentile Rank of SCT Scores— 2014 cohort	*Lowest quartile for SCT scores (i.e., lowest 25% of SCT scores) Count (%)*	44 (77.19)	13 (22.81)	228	χ^2 (3,228) = 12.14 p = 0.007
	2nd quartile for SCT scores (25th percentile rank to 50th percentile rank) Count (%)	38 (66.67)	19 (33.33)		
	3rd quartile for SCT scores (50th percentile rank to 75th percentile rank) Count (%)	30 (52.63)	27 (47.37)		
	Top quartile for SCT scores (i.e., highest 25% of SCT scores) Count (%)	28 (49.12)	29 (50.87)		
Quartile/ Percentile Rank of SCT Scores— 2015 cohort	*Lowest quartile for SCT scores (i.e., lowest 25% of SCT scores) Count (%)*	32 (58.18)	23 (41.82)	220	χ^2 (3,220) = 13.45 p = 0.004
	2nd quartile for SCT scores (25th percentile rank to 50th percentile rank) Count (%)	27 (49.09)	28 (50.91)		
	3rd quartile for SCT scores (50th percentile rank to 75th percentile rank) Count (%)	21 (38.18)	34 (61.82)		
	Top quartile for SCT scores (i.e., highest 25% of SCT scores) Count (%)	14 (25.45)	41 (74.55)		

RESPONSE PATTERN IN SCRIPT CONCORDANCE TEST

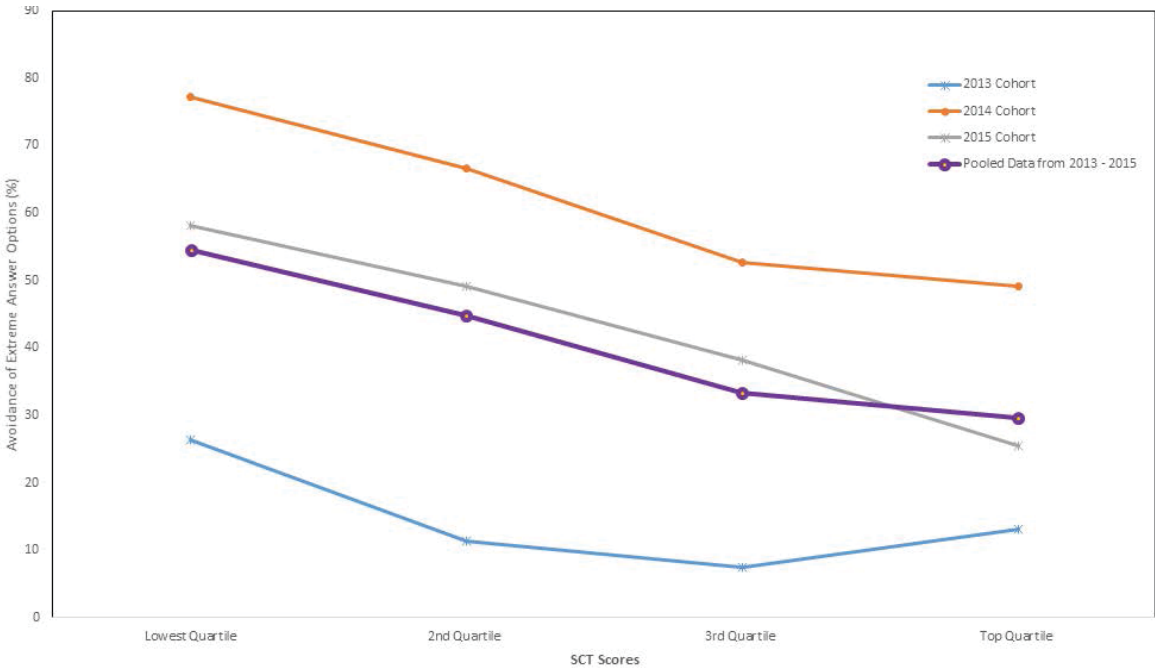


Figure 2. Percentage avoidance of extreme-response options in SCT by candidates' overall performance by quartile in SCT scores.

Discussion

Data from our study shows a significant negative association between overall SCT scores and the propensity to avoid the extreme-response options. This negative association suggests that candidates who tend to avoid extreme-response options do not achieve inflation of their SCT scores, in contrast to the findings from Lineberry, Kreiter and Bordage (2013). A further follow-up study using post-hoc simulation and rescoring of SCT data will provide more evidence on the actual impact of extreme-response options avoidance on candidates' overall SCT scores.

The response pattern, that is, propensity to avoid extreme options, of the students whose SCT scores were in the lowest quartile, could be due to a test-taking strategy or avoidance of the extreme-response options simply because they were not confident about the likelihood of a diagnosis or management plan (due to poor command of basic clinical science knowledge). Such avoidance obviously did not advantage them in terms of getting higher scores.

The aforementioned findings could be a result of some pre-emptive strategies in the study context. Apart from fulfilling the usual blueprinting in ensuring a sufficient spread of clinical scenarios for representativeness of item sampling for each SCT paper, items with roughly equal number of full marks in each option across the five response options are selected from the SCT-item pool. In other words, to mitigate the impact of any test-taking strategies that may have been adopted by students, we select SCT items with modal answers from the expert reference panel that cover the full 5-point

Likert scale response options. Students should not be advantaged or disadvantaged by selecting predominantly the "-1", "0" or "+1" response options and avoiding the extreme options of "-2" and "+2". Student performance on SCT tests will then more likely reflect student expertise in clinical reasoning rather than expertise in test-taking behaviour, or confidence in reaching a definitive decision.

While the data for this study only came from one medical school, the study sample was reasonably large (n = 660) and included six cohorts of students. The findings and resulting recommendations related to construction of SCT items should be generalisable to other settings. A limitation of this study is the pure quantitative method used in the analysis. A think-aloud protocol would have been useful to analyse the actual reasons behind the candidates' avoidance of the extreme-response options in SCTs.

Therefore, another study is underway to look at the underlying reasons for candidates avoiding the extreme responses. A focus group discussion and think-aloud analysis will look deeper into what is in the students' mind when they choose to avoid the extreme-response options in SCT, i.e., whether this avoidance behaviour is due to lack of confidence in their command of clinical science knowledge for clinical reasoning and decision making, or it is a conscious test-taking strategy employed by the students.

Before conclusive recommendations can be made, further work to investigate the issue of potential threats to validity of SCT scores are crucial, particularly using empirical data from other medical schools using SCT as an assessment modality. A simulation study through post-hoc rescoring of current SCT data set (as briefly mentioned before) will be conducted in this study context to further investigate the extent of score inflation in SCT as a result of complete avoidance of extreme-response options (by recoding "-2" to "-1"; "+2" to "+1") or as a result of only choosing "0" as the answer to all items which were performed by other colleagues (Lineberry et al., 2013; See et al., 2014).

Conclusions

Students whose SCT scores are in the lowest quartile seem more likely to avoid the extreme-response options in answering SCT questions.

Developing good-quality SCT questions is not easy. As with all other assessment modalities, careful planning and development of SCT items, along with necessary quality assurance and quality monitoring mechanisms, are crucial to mitigate possible threats to the validity of SCT scores. Acknowledging the vulnerability of SCT scores to possible validity threats due to the format of SCT response options and the characteristics of aggregate partial-credit scoring models is crucial. As demonstrated by the study findings, careful construction and selection of items that can be built into the SCT development procedures may be helpful to mitigate some of the plausible threats to validity of SCT scores. Particular care should be taken to develop SCT items that could attract the full range of the 5 response options available for student answer choice. In other words, the additional pieces of new information should result in the consideration of "-2" and "+2" as well as "-1", "0" and "+1" options.

References

Boulouffe, C., Doucet, B., Muschart, X., Charlin, B., & Vanpee, D. (2014). Assessing clinical reasoning using a script concordance test with electrocardiogram in an emergency medicine clerkship rotation. *Emergency Medicine Journal*, *31*(4), 313–316. doi:10.1136/emermed-2012-201737

Brazeau-Lamontagne, L., Charlin, B., Gagnon, R., Samson, L., & van der Vleuten, C. (2004). Measurement of perception and interpretation skills during radiology training: Utility of the script concordance approach. *Medical Teacher*, *26*(4), 326–332. doi:10.1080/01421590410001679000

Carrière, B. (2009). Assessing clinical reasoning in pediatric emergency medicine: Validity evidence for a script concordance test. *Annals of Emergency Medicine*, *53*(5), 647–652. doi:10.1016/j.annemergmed.2008.07.024

Chang, T. P., Kessler, D., McAninch, B., Fein, D. M., Scherzer, D. J., Seelbach, E., . . . Education, N. (2014). Script concordance testing: Assessing residents' clinical decision-making skills for infant lumbar punctures. *Academic Medicine*, *89*(1), 128–135. doi:10.1097/ACM.0000000000000059

Charlin, B., Roy, L., Brailovsky, C., Goulet, F., & van der Vleuten, C. (2000). The script concordance test: A tool to assess the reflective clinician. *Teaching and Learning in Medicine*, *12*(4), 189–195. doi:10.1207/S15328015TLM1204_5

Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment*, *14*(2), 142–155. doi:10.1111/j.1468-2389.2006.00340.x

Dawson, T., Comer, L., Kossick, M. A., & Neubrander, J. (2014). Can script concordance testing be used in nursing education to accurately assess clinical reasoning skills? *Journal of Nursing Education*, *53*(5), 281–286. doi:10.3928/0148434-20140321-03

Dory, V., Gagnon, R., Vanpee, D., & Charlin, B. (2012). How to construct and implement script concordance tests: Insights from a systematic review. *Medical Education*, *46*(6), 552–563. doi:10.1111/j.1365-2923.2011.04211.x

Drolet, P. (2015). Assessing clinical reasoning in anesthesiology: Making the case for the script concordance test. *Anaesthesia Critical Care & Pain Medicine*, *34*(1), 5–7. doi:10.1016/j.accpm.2015.01.003

Ducos, G., Lejus, C., Sztark, F., Nathan, N., Fourcade, O., Tack, I., . . . Minville, V. (2015). The script concordance test in anesthesiology: Validation of a new tool for assessing clinical reasoning. *Anaesthesia Critical Care & Pain Medicine*, *34*(1), 11–15. doi:10.1016/j.accpm.2014.11.001

Duggan, P., & Charlin, B. (2012). Summative assessment of 5th year medical students' clinical reasoning by script concordance test: Requirements and challenges. *BMC Medical Education*, *12*(1), 29–29. doi:10.1186/1472-6920-12-29

Foucault, A., Dubé, S., Fernandez, N., Gagnon, R., & Charlin, B. (2015). Learning medical professionalism with the online concordance-of-judgment learning tool (CJLT): A pilot study. *Medical Teacher*, *37*(10), 955.

Irfannuddin, I. (2009). Knowledge and critical thinking skills increase clinical reasoning ability in urogenital disorders: A Universitas Sriwijaya medical faculty experience. *Medical Journal of Indonesia*, *18*(1), 53–59. doi:10.13181/mji.v18i1.341

Kow, N., Walters, M. D., Karram, M. M., Sarsotti, C. J., & Jelovsek, J. E. (2014). Assessing intraoperative judgment using script concordance testing through the gynecology continuum of practice. *Medical Teacher*, *36*(8), 724–729. doi:10.3109/0142159X.2014.910297

Lambert, C., Gagnon, R., Nguyen, D., & Charlin, B. (2009). The script concordance test in radiation oncology: Validation study of a new tool to assess clinical reasoning. *Radiation Oncology*, *4*(1), 7. doi:10.1186/1748-717X-4-7

Lineberry, M., Kreiter, C. D., & Bordage, G. (2013). Threats to validity in the use and interpretation of script concordance test scores. *Medical Education*, *47*(12), 1175–1183. doi:10.1111/medu.12283

Lubarsky, S., van der Vleuten, C. P. M., Charlin, B., Chalk, C., & Cook, D. A. (2011). Script concordance testing: A review of published validity evidence. *Medical Education*, *45*(4), 329–338. doi:10.1111/j.1365-2923.2010.03863.x

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *The Journal of Applied Psychology*, *96*(2), 327–336. doi:10.1037/a0021983

Nouh, T., Boutros, M., Gagnon, R., Reid, S., Leslie, K., Pace, D., . . . Meterissian, S. H. (2012). The script concordance test as a measure of clinical reasoning: A national validation study. *American Journal of Surgery*, *203*(4), 530–534. doi:10.1016/j.amjsurg.2011.11.006

Palmer, E. J., Duggan, P., Devitt, P. G., & Russell, R. (2010). The modified essay question: Its exit from the exit examination? *Medical Teacher*, *32*(7), e300–e307. doi:10.3109/0142159X.2010.488705

See, K. C., Tan, K. L., & Lim, T. K. (2014). The script concordance test for clinical reasoning: Re-examining its utility and potential weakness. *Medical Education*, *48*(11), 1069–1077. doi:10.1111/medu.12514

Tan, K., Tan, N., Kandiah, N., Samarasekera, D., & Ponnamperuma, G. (2011). A script concordance test for neurological localization and emergencies. *Neurology*, *76*(9), A268–A268.

Tsai, T.-C., Chen, D.-F., & Lei, S.-M. (2012). The ethics script concordance test in assessing ethical reasoning. *Medical Education*, *46*(5), 527–527. doi:10.1111/j.1365-2923.2012.04252.x

Wan, S. H. (2014). Using script concordance testing (SCT) to assess clinical reasoning: The progression from novice to practising general practitioner. *Medical Education*, *48*(2), 6.

Wan, S. H. (2015). Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine. *Hong Kong Medical Journal*, *21*(5), 455–461. doi:10.12809/hkmj154572

Synopsis of Chapter 3

Chapter 3 reported evidence based on the data collected in the study context, in actual summative examination settings, that the students in the lowest quartile of SCT scores were more likely to avoid the extreme response options in answering SCT questions. The alleged test-taking strategy could potentially threaten the response process validity of SCT scores. Findings from this study indicated that for high stakes summative examinations, careful selection of SCT items with expert reference panel's modal answers covering the full 5-point response options appears necessary, to mitigate this test taking strategy and improve the validity of the test scores.

Chapter 4: Improving the validity of script concordance testing by optimising and balancing items

This chapter contains the paper titled “Improving the validity of script concordance testing by optimising and balancing items” published in Medical Education (ERA Journal) 2018;52(3):336-46.

Statement of Contribution by Others: please refer to Appendix 3

Foreword

With evidence from the study reported in Chapter 3 that lower performing students more commonly avoid the extreme response options in SCT, a follow-up post-hoc simulation study tested the hypothesis that test-wise students’ SCT scores were inflated through deliberate avoidance of extreme responses; and deliberate selection of only the neutral responses. This study also investigated whether better optimisation and balancing of the items in the SCT paper could help to mitigate the possible score inflation from test-wise answering strategies; therefore mitigating some of the response process validity threats to the SCT test scores.

Approval granted from Journal Editor for inclusion in the Thesis (Appendix 9).

Improving the validity of script concordance testing by optimising and balancing items

Michael SH Wan,[1] Elina Tor[1] & Judith Nicky Hudson[2]

BACKGROUND A script concordance test (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 students' 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 an SCT.

METHODS Using the paired sample *t*-test, the actual average student scores for 10 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 with those from 2013–2016, 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 simulation 2 (46.7% and 23.9% respectively).

CONCLUSIONS When constructing SCT examinations, apart from the rigorous pre-examination optimisation, it is desirable to achieve a balance between items that attract extreme responses and those that attract 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.

Medical Education 2018: 52: 336–346
doi: 10.1111/medu.13495

[1]School of Medicine, University of Notre Dame, Sydney, New South Wales, Australia
[2]Adelaide Medical School, University of Adelaide, Adelaide, South Australia, Australia

Correspondence: Michael SH Wan, Medical Education Unit, The University of Notre Dame, Sydney, 160 Oxford Street, Darlinghurst. New South Wales 2010, Australia. Tel: 028 204 4479; E-mail: michael.wan@nd.edu.au

INTRODUCTION

A script concordance test (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 five-point Likert scale. In an 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', for more likely a '+1', for neither less nor more likely a '0', for less likely a '−1', and for much less likely a '−2'. 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 questions, in thinking of the diagnosis of carcinoma of the colon, the student will not consider the additional information of a normal blood glucose or thyroid stimulating hormone (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.

To score the SCT items, the student's selection is compared with 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 panelists' 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]

The SCT has been used in undergraduate medical school examinations as well as in postgraduate fellowship training. Successful implementation of

Table 1 Sample script concordance test questions

Clinical scenario

A 45-year-old woman presents to the general practitioner clinic with weight loss of 5 kg in 2 months. She has no significant past medical history.

	If you were thinking of . . .	*and then you find that . . .*	*this hypothesis becomes . . .*
Q1.	Diabetes mellitus	Normal fasting blood sugar	A B C D E −2 −1 0 +1 +2
Q2.	Graves' disease	Normal TSH level	A B C D E −2 −1 0 +1 +2
Q3.	Carcinoma of the colon	A normal digital rectal examination	A B C D E −2 −1 0 +1 +2

−2: much less likely; −1: less likely; 0: neither more nor less likely; +1: more likely; +2: much more likely. TSH = thyroid stimulating hormone

Table 2 Formula to calculate the weighted scores in script concordance testing

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

script concordance testing has been documented in medicine, surgery, paediatrics, emergency medicine, anaesthesia, psychiatry and ethics.[5–10] Script concordance testing 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 script concordance testing has been shown by various studies demonstrating progression of SCT scores from undergraduate medical students to postgraduate fellows in training.[8,15,18–20]

However, a recent study has suggested that the aggregate partial credit scoring method used in SCTs could be subjected to validity threats.[17] Lineberry et al. showed that 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 selected SCT 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' for 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 judgement tests, which also use a partial aggregate scoring approach.[21,22]

In another study using two sets of 96-item SCTs on pulmonary and critical care for postgraduate trainees, simply avoiding extreme answers boosted the Z-scores of the lowest 10 scorers on both SCT sets by ≥1 SD.[14] 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.[23] 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 SCTs would result in an increase in the average SCT scores for the whole cohort 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 SCTs in the summative examinations for Year 3 and Year 4 clinical students in the 4-year graduate-entry medical programme 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 and gynaecology in Year 3, and 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 middle of the year to each year group as a formative examination.

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 pass or fail score of each SCT examination we used the expert reference panel's mean minus 4 standard deviations (SDs) as the cut-score.[7]

Test optimisation processes

After each panel's scoring, a test optimisation process is conducted where questions with (i) bimodal, (ii) uniform divergence and (iii) discrete outlier responses from the panel are discarded, reducing expert disagreement in the answers. The remaining items are all with the (iv) 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 Fig. 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.[24] 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) on 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.

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 a 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.

Data analysis

Simulation 1. Simulation 1 involved *post-hoc* recoding of all '−2' responses to '−1' and '+2' responses to '+1'. Actual average SCT scores and the average

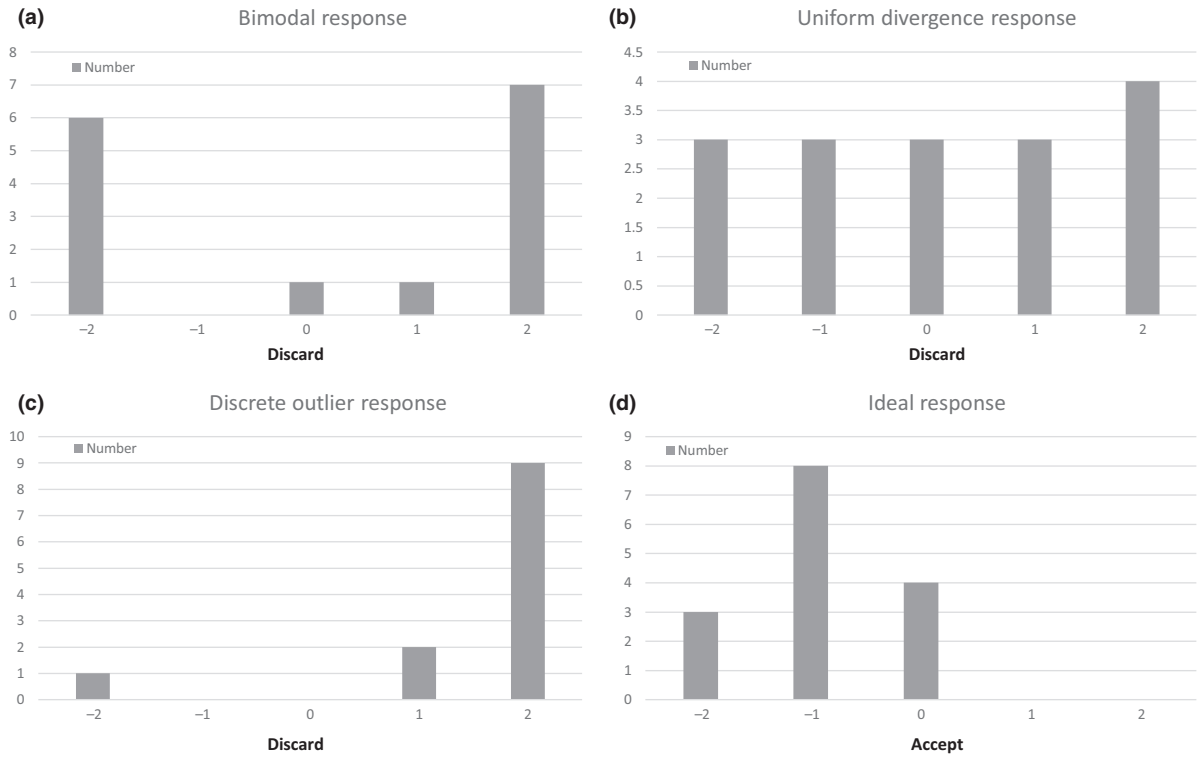


Figure 1 Expert panel responses to questions in a script concordance test. Bimodal response (a), uniform divergence response (b), discrete outlier response (c) and ideal response (d)

Table 3 Sample 2016 script concordance test 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				

All full marks (modal responses) highlighted in bold.

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 to 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 to 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.[14,17] 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 with post-intervention scores (2013–2016). The paired sample

Table 4 Example of recoding involved in simulations 1 and 2

Script concordance test 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

t-test was used for statistical analysis of the comparisons (IBM, SPSS, Inc., Chicago, IL, USA; 24).

Ethical 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 programme for the Bachelor of Medicine, Bachelor of Surgery (MBBS) course and a total of 10 cohorts of 120 students each have been examined during this time. Since 2013, with the 'balanced approach', there has been 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-item examinations for the 5 years is shown in Table 5.

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(a–e) shows the original and simulated scores from 2012 to 2016, respectively, by quartiles. In 2012 (Fig. 2a), before the 'balanced approach' was introduced, the avoidance of extreme answer options did not seem to have a huge impact on SCT scores, as indicated by the closeness between the two line-graphs for original and *post-hoc* simulated recoded scores. Figure 2(a) also shows that in 2012, students in the bottom quartile may

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

	Number of SCT items with modal answer: a (−2)	b (−1)	c (0)	d (+1)	e (+2)	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
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.

have achieved 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.

As a result of 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 eight 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 with the original average SCT score (50.15%) for the bottom 10% of the cohort. By 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.

These comparisons are represented pictorially in Appendix S1 to highlight the findings: for the bottom 10% of students in the 2012 Year 3 cohort, avoidance of extreme answer options has resulted in a significant increase in SCT score, whereas in 2013–2016 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 indicate 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 of 13.7–35% (Table 7).

DISCUSSION

This study has investigated a possible intervention that may mitigate one of the threats to SCT validity

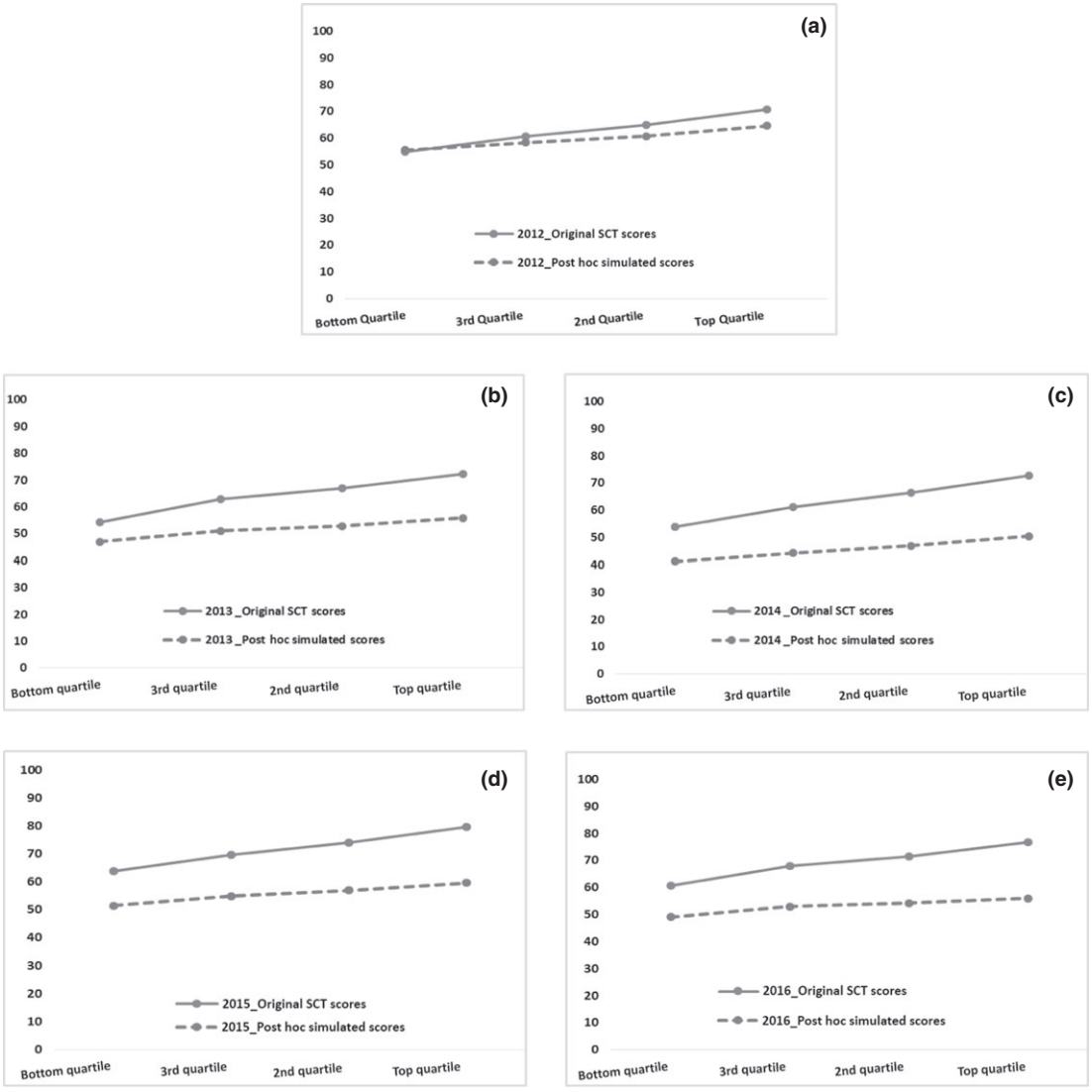


Figure 2 Original and simulated scores from 2012–2016 according to the performance ability in quartiles of the students

as a result of 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 or her test results. The current study has shown that SCT test optimisation processes, such as balancing the distribution of the expert reference panel's modal answers for items in SCTs across the whole continuum of the Likert response scale, and controlling for other conceptual or logical flaws in partial aggregate scoring used for the conventional SCTs, 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 script concordance testing, examinees who potentially choose to deliberately avoid extreme answer options, 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.[14,17] The analysis of baseline data for all students in 2012 prior to the

Table 6 Effect of deliberate avoidance of extreme answer options on script concordance test (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	−2.75	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.

Table 7 Simulated scores if students chose '0' for all items compared with actual cohort mean scores in Year 3 and 4 from 2012 to 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 2013–2016	**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.

implementation of a 'balanced approach' also shows that there was an interaction effect between examinees' ability (using score quartiles as a proxy) and the prevalence and extent of score 'inflation' through deliberate avoidance of extreme answer options (Fig. 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).[23] 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 the expert reference panel's response pattern in SCT items (the 'balanced approach'), could remove the potential 'score inflation' previously described.[17]

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 into the design of each SCT paper, there is always a somewhat balanced distribution of median and extreme options in the expert panel's modal answers that will attract a full mark. Students should be urged to respond according to their knowledge, understanding and reasoning, using 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 the expert reference panel is certainly a useful validity feature that can be built into 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 script concordance testing, particularly in relation to the logical inconsistency of the answer responses and the accuracy of the expert panel's answers compared with the evidence-based likelihood ratios.[16,17] The pre-existing test optimisation procedures adopted in the study context have addressed the concern over the faulty logic of aggregated scoring in SCTs.[17] The selection of only the 'ideal' responses from the panel and reducing the experts' disagreements 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 'bimodal' and 'discrete outlier' response patterns 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 regarding 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 because of 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 students' 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.[25] This process may help address the concern raised by Kreiter (2012)[26] 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. Indeed, in a response to this, Lubarsky et al.[27] 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. 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 of plausible validity threats to SCT scores as a result of 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 SCTs. On the contrary, we recommend that the hypotheses and conclusion derived from this study be further tested in other medical education settings.

Contributions: MSHW: contributed to the acquisition, analysis and interpretation of the data, writing up and continual revision of the draft manuscript. ET: contributed to the statistical analysis of the raw data and data interpretation, and continual revision of manuscript content. NH: contributed to the interpretation of the data, editing of the manuscript drafts and continual revision of the paper. The final version of the paper has been approved by all authors, who agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.
Acknowledgements: None.
Funding: None.
Competing interests: None.
Ethical approval: Approval has been given by the University of Notre Dame Human Research Ethics Committee, Reference Number 016126S.

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. *Teach Learn Med* 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 Med J* 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. *Med Teach* 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? *Med Educ* 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. *Anaesth Crit Care Pain Med* 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 Med Educ* 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. *Acad Psychiatry* 2017;**41** (1):86–90.
9 Nouh T, Boutros M, Gagnon R, et al. The script concordance test as a measure of clinical reasoning: a national validation study. *Am J Surg* 2012;**203** (4):530–4.
10 Tsai T-C, Chen D-F, Lei S-M. The ethics script concordance test in assessing ethical reasoning. *Med Educ* 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. *Clin Exp Optom* 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**:e332–3.
13 Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script concordance testing: a review of

published validity evidence. *Med Educ* 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. *Med Educ* 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. Abstracts of the 11th Asia Pacific Medical Education Conference, Singapore, *Med Educ* 2014;**48** (Suppl. 2):6.
16 Ahmadi S-F, Khoshkish S, Soltani-Arabshahi K, Hafezi-Moghadam P, Zahmatkesh G, Heidari P, Baba-Beigloo D, Baradaran H, Lotfipour S. Challenging script concordance test reference standard by evidence: do judgments by emergency medicine consultants agree with likelihood ratios? *Int J Emerg Med* 2014;**7**:34.
17 Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. *Med Educ* 2013;**47** (12):1175–83.
18 Ducos G, Lejus C, Sztark F, Nathan N, Fourcade O, Tack I, Asehnoune K, Kurrek M, Charlin B, Minville V. The Script Concordance Test in anesthesiology: validation of a new tool for assessing clinical reasoning. *Anaesth Crit Care Pain Med* 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. *Acad Pediatr* 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. *Acad Med* 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. *Int J Selection Assess* 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. *J Appl Psychol* 2011;**96** (2):327–36.
23 Wan SH, Duggan P, Tor E, Hudson JN. Association between candidate total scores and response pattern in script concordance testing of medical students. *Focus Health Prof Educ* 2017;**18** (2):26–35.
24 Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC Med Inform Decis Mak* 2008;**8** (1):18.
25 Power A, Lemay J-F, Cooke S. Justify your answer: the role of written think aloud in script concordance testing. *Teach Learn Med* 2017;**29** (1):59–67.
26 Kreiter CD. Commentary: the response process validity of a script concordance test item. *Adv Health Sci Educ* 2012;**17** (1):7–9.
27 Lubarsky S, Gagnon R, Charlin B. Script concordance test item response process: the argument for probability versus typicality. *Adv Health Sci Educ* 2012;**17** (1):11–3.

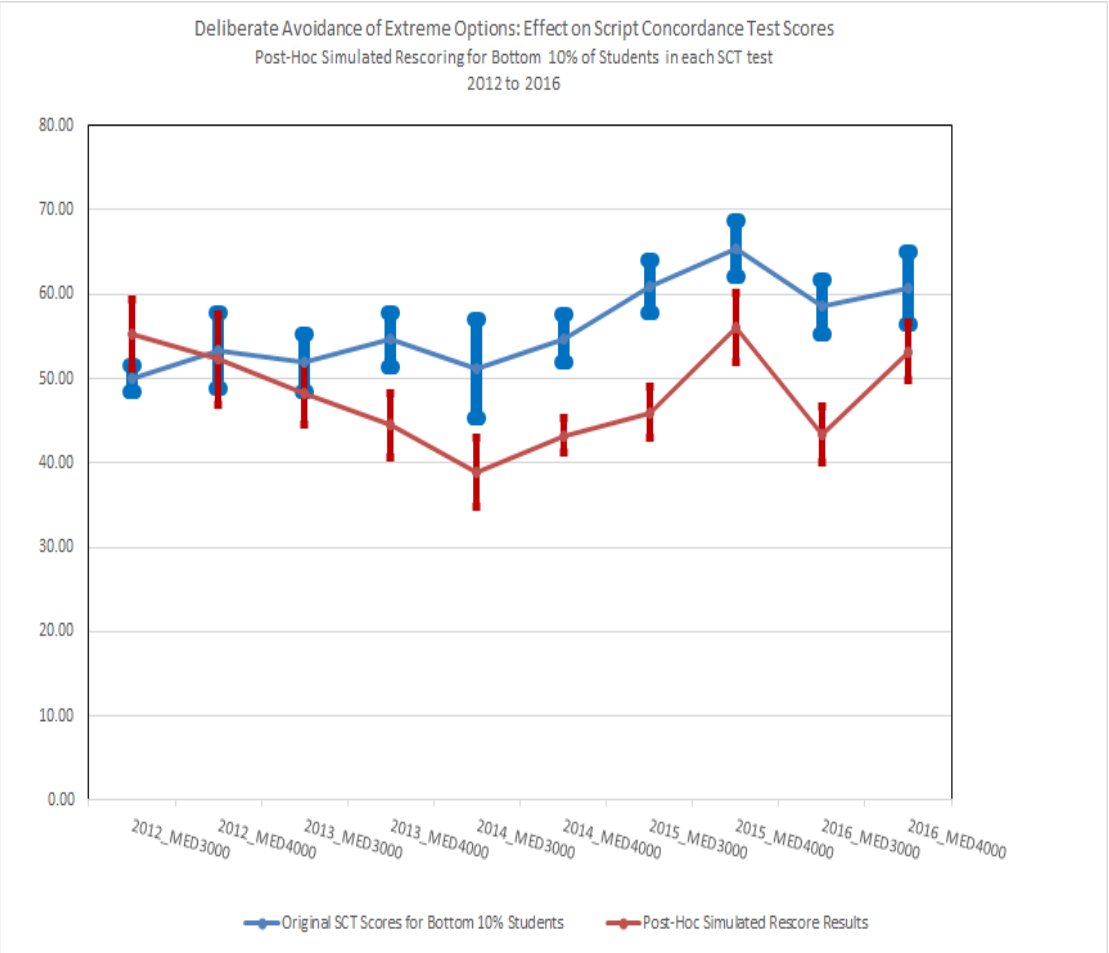
SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Effect on script concordance test (SCT) scores by deliberate avoidance of extreme options for bottom 10% of students (2012–2016).

Received 27 June 2017; editorial comments to author 18 August 2017, 17 October 2017; accepted for publication 19 October 2017

Appendix S1. Effect on Script Concordance Test (SCT) scores by deliberate avoidance of extreme options for bottom 10% of students (2012-2016).



Synopsis of Chapter 4

Chapter 4 presents the findings from a study in the research context comparing the actual and simulated scores before and after the implementation of the ‘test optimisation and balancing’ approach in constructing high stakes SCT examinations. This study aimed to investigate the effect of possible score inflation due to the test taking strategy of students. The results indicated that a balanced distribution of extreme and median options in the modal responses by the expert reference panel is an important validity feature that can be built into the SCT test development process (i.e. as the final step in the routine SCT test optimisation process). This further supports the response process validity of SCT scores.

Chapter 5: Construct validity of Script Concordance Testing scores: progression from medical students to general practitioners

This chapter contains the paper titled "Construct validity of Script Concordance Testing scores: progression from medical students to general practitioners" published in IJME (ERA Journal) 2019;10:174.

Statement of Contribution by Others: please refer to Appendix 4

Foreword

In response to the paucity of empirical studies on the construct validity for SCT scores, this chapter reported a study comparing the SCT scores of Year 3 and Year 4 clinical year medical students, junior general practice registrars and experienced practising general practitioners (GPs). The progression of scores, noted in the study context, provided further evidence of the construct validity of the SCT. This continual progression in SCT scores from medical students; to junior registrars and GPs had not been reported in published literature so far.

International Journal of Medical Education. 2019;10:174-179
ISSN: 2042-6372
DOI: 10.5116/ijme.5d76.1eee

Construct validity of script concordance testing: progression of scores from novices to experienced clinicians

Michael Siu Hong Wan[1], Elina Tor[1], Judith N. Hudson[2]

[1]School of Medicine, University of Notre Dame, Australia
[2]Faculty of Health and Medical Sciences, University of Adelaide, Australia

Correspondence: Michael Siu Hong Wan, 160 Oxford Street, Darlinghurst, NSW 2010, Australia
Email: michael.wan@nd.edu.au

Accepted: September 09, 2019

Abstract

Objective: To investigate the construct validity of Script Concordance Testing (SCT) scores as a measure of the clinical reasoning ability of medical students and practising General Practitioners with different levels of clinical experience.
Methods: Part I involved a cross-sectional study, where 105 medical students, 19 junior registrars and 13 experienced General Practitioners completed the same set of SCT questions, and their mean scores were compared using one-way ANOVA. In Part II, pooled and matched SCT scores for 5 cohorts of students (2012 to 2017) in Year 3 (N=584) and Year 4 (N=598) were retrospectively analysed for evidence of significant progression.
Results: A significant main effect of clinical experience was observed [$F_{(2, 136)}$=6.215, p=0.003]. The mean SCT score for General Practitioners (M=70.39, SD=4.41, N=13) was significantly higher (p=0.011) than that of students (M = 64.90, SD = 6.30, N=105). Year 4 students (M=68.90, SD= 7.79, N=584) scored a significantly higher mean score [$t_{(552)}$=12.78, p<0.001] than Year 3 students (M = 64.03, SD=7.98, N=598).
Conclusions: The findings that candidate scores increased with increasing level of clinical experience add to current evidence in the international literature in support of the construct validity of Script Concordance Testing. Prospective longitudinal studies with larger sample sizes are recommended to further test and build confidence in the construct validity of SCT scores.
Keywords: Script Concordance Testing, validity, assessment, clinical reasoning

Introduction

Since 2009, Script Concordance Testing (SCT) has been used to assess higher-order clinical reasoning and data interpretation skills in the context of uncertainty, at both undergraduate and postgraduate medical education levels.[1] It was designed to probe one key signpost along an accepted theoretical pathway of clinical reasoning under uncertainty.[2] In each SCT, candidates are presented with a clinical scenario, followed by a new piece of information. The candidates are then asked to assess whether this additional piece of information increases or decreases the probability of the suggested provisional diagnosis or increases or decreases the appropriateness of a proposed investigation or management option. In the classical scoring of SCT, the candidate's decision is compared to that of a reference panel of experts in the field and a weighted partial scoring system with a 5-point Likert scale is applied.[3] Since its development, the SCT format has been used in assessment across many medical disciplines, including Medicine, Surgery, Psychiatry, Paediatrics, Dentistry and more recently, Medical Ethics.[4-12]

As for all educational assessments, SCT use as summative assessment in Medicine requires evidence to support the appropriateness and meaningfulness of interpretation and use of the results.[13] Over the past few years a number of studies in the international literature have addressed some issues on the validity of SCT scores.[9,14] However, there is a relative paucity of evidence demonstrating that SCT scores are a measure that can discriminate between the reasoning skills of medical practitioners at different stages in their medical career – i.e. from medical students, to junior doctors, to experienced doctors. This is an important piece of evidence for the overall construct validity of SCT scores, an issue which this study aims to address. In general, construct validity is the degree to which an instrument measures the construct it is intended to measure.[15,16] In the context of Script Concordance Testing,

according to key developers of this assessment format, the construct validity of scores from script concordance testing depends on the inference that candidates with more evolved illness scripts interpret data and make judgments in uncertain situations that increasingly concord with those of experienced clinicians given the same clinical scenarios.[3] The tendency for SCT scores to consistently increase with increasing level of training has been reported as empirical evidence supporting the validity of this inference.[17]

The progression of clinical reasoning capability, as measured by SCT in post-graduate medical education settings, has been reported in previous studies. In 2009 Lubarsky showed that Neurology trainees' SCT scores improved as they progressed through the post-graduate training program. This evidence of progression of SCT scores supported the construct validity of SCT in this setting.[1] There is also evidence of progression of clinical reasoning during residency emergency training in Paediatrics.[14] Kazour examined interns (junior doctors in the first post-graduate year) using a set of 100 SCT questions in Psychiatry and found significant improvement in the interns' scores between the beginning and the end of their rotation.[8] A further study used SCT scenarios to assess the reasoning skills of paediatric residents and neonatal-perinatal medicine fellows (qualified specialists), and reported a significant difference between all training levels from Post-graduate Year 1 (PGY-1) to PGY-3 and between PGY-3 and fellows, with improvement of scores observed for each progressive level of medical training.[18] More recently, Subra administered an SCT assessment to post-graduate students in general practice and showed progression of clinical reasoning throughout the 3 years of training pathway especially in the first 18-months.[19] However, there is an apparent gap in the literature, specifically in relation to empirical evidence of progression in clinical reasoning skills for medical students in undergraduate medical education. Furthermore, studies comparing the clinical reasoning capability of medical students and practising clinicians, using the same set of SCT items, are lacking. This study aimed to address these gaps by seeking evidence of progression of medical students' SCT scores through the two senior clinical years, and evidence of higher scores for experienced clinicians and post-graduate trainees when compared with those of senior medical students (novices), on the same set of SCT questions. Progression in performance on SCTs, i.e. tendency for SCT scores to consistently increase with increasing level of training and experience, should provide further support for the hypothesis that SCT scores are a valid measure of clinical reasoning ability in Medicine.

The setting for the current study was a medical school in Australia with a four-year graduate-entry medical program. The School has been using SCT questions in Year 3 and Year 4 summative assessments of the program since 2010. The aim of the study was to investigate the construct validity of SCT scores as a measure of clinical reasoning ability of senior medical students and practising clinicians of differing experience in general medical practice (family medicine). Specifically, this study sought to test the following hypotheses for the construct validity of SCT scores:

1. There is a significant progression in SCT scores from senior medical students, junior registrars, to experienced general practitioners (GPs), using the same set of SCT questions (Part I)
2. There is a significant improvement in SCT scores when students progress from Year 3 to Year 4 within the clinical phase of their undergraduate medical program, as measured by retrospective analysis of pooled and matched (same cohort) SCT assessment scores for 5 cohorts of medical students (Part II)

Methods

Study design and participants

This study was a two-pronged practitioner inquiry within one medical school context in Australia.[20] Part I of the study involved a cross-sectional study design. Three groups of participants took part in the study: final year medical students (N=105) completed the 40 items SCT as part of their invigilated written summative examination in October 2015; registrars in general practice training (N=19); and practising General Practitioners (N=13), completed the same SCT paper in January 2016. The registrars who were junior doctors with less than 4 years post-graduation clinical experience, were recruited via general email invitations distributed to School alumni. The General Practitioners (GPs) participants, who had at least 5 years of post-fellowship practice experience in General Practice, were also part-time Problem Based Learning (PBL) tutors at the School. Both groups of graduated doctors were volunteer participants in the study.

The expert reference panel (N = 17) comprised specialists in relevant disciplines who had provided answers to the SCT questions in the written paper, and their responses to each SCT item were used as the basis for the scoring of the responses by participants in the three study groups, using the classical weighted aggregate partial scoring approach.[3]

Part II of the study involved a retrospective analysis of pooled and matched summative SCT assessment scores for 5 cohorts of medical students (2012 to 2017). Expert reference panels (N=13-18) comprised specialists in the relevant disciplines and their responses were used as the basis for the scoring using the same approach as in Part I.

Ethics approval for the study was obtained from the University's Human Research and Ethics Committee (#018161S). To ensure the confidentiality and anonymity of the participants, all data were de-identified prior to the commencement of data analysis.

Data collection

In Part I of the study, a set of 40 SCT questions, based on 15 case scenarios covering the disciplines of Medicine, Surgery, Paediatrics, Obstetrics & Gynaecology, Psychiatry and

General Practice, were developed to assess clinical reasoning according to the assessment blueprint of the medical programme. Each SCT question was reviewed by discipline-specific experts and the assessment academics at the School to ensure content validity. The usual format for construction of SCT items was employed.[12,21] Special attention was made to ensure that there was a balance between items that attract extreme responses and those that attract median response options. As previously reported in the literature, careful balancing of item response options aimed to minimise the threats to validity by test-wise students who may try to game the examination, or the lowest quartile students trying to avoid extreme option answers.[22-24] The set of 40 SCT questions were given to the 3 participant groups described above. Prior to this end-of-year summative examination (October 2015), all medical students had sat for a formative mid-year SCT examination and a practice online SCT quiz. The GP registrars, and experienced GPs completed the same set of SCT items in January 2016. Junior registrars and experienced GP study participants were given a detailed explanation of the structure and scoring of SCT as well as a sample set of SCT questions before they were asked to answer the set of SCT questions. After providing consent to participate, the junior registrars attempted the SCT online, using a survey template where all answers were collected anonymously. The experienced GP participants attempted the same set of SCT questions on campus under invigilation, using a paper-based format similar to the medical students.

Part II of the study involved retrospective analysis of the summative SCT scores for five cohorts of medical students in their clinical years from 2012 to 2017 inclusive. As part of the invigilated written end-of-year summative assessment at the School, all students completed a set of 40 SCT questions in their penultimate (Year 3) and final (Year 4) clinical year. As in Part I of the study, SCT scoring was based on the classical aggregated partial scoring method, with a full mark being awarded for concordance with the majority of the expert panel and a partial weighted score for concordance with the minority of the panel.[12]

Data analysis

A one-way between-subjects ANOVA was conducted on the data from the first part of this study, to compare the difference in mean SCT scores obtained by senior medical students, GP registrars and practising GPs. For the comparison, $p < .05$ was considered statistically significant. In the second part of the study, pooled and matched SCT scores in Year 3 and Year 4 for individual students from five cohorts (2012 to 2017), were analysed for evidence of significant progression, or the lack thereof, using a repeated measure t-test. Each student's Year 3 SCT score was paired and matched with their respective Year 4 scores when the student had progressed to the final year of their MBBS/MD program. The SPSS statistical package version 25 (IBM Corp., Armonk, NY) was used for the statistical analysis in both parts of the study.

Results

The Cronbach's alpha value for scores from each SCT paper was in the range of 0.62 to 0.86 (2012-2017), providing evidence of acceptable reliability (i.e. internal consistency) of the SCT scores. Part I of the study indicated a significant main effect of clinical experience on performance in the SCT, at the $p<0.05$ level for the three stages of medical career i.e. medical students, junior GP registrars, and, experienced GPs [$F_{(2,136)}=6.215$, $p=0.003$]. The effect size (Eta squared, $\eta^2=0.084$) was moderate, based on Cohen's guidelines (small effect size: $\eta^2= 0.01$; medium effect size - $\eta^2=0.06$; large effect size - $\eta^2= 0.14$).[25,26]

Post hoc comparisons using the Bonferroni method indicated that the mean SCT score for experienced GPs (M = 70.39, SD=4.41, N=13) was significantly higher (p=0.011) than the mean SCT score of medical students (M= 64.90, SD = 6.30, N=105). However, the mean SCT score for junior GP registrars (M=68.36, SD=7.20, N=19) did not differ significantly from the mean SCT scores of medical students (p= 0.069) and the experienced GPs (p=1.000). The expert panel's (N=17) average score was 79.40% (SD=10.8). The results are represented in the box plot (Figure 1).

Part II of the study compared pooled and matched data for five cohorts (2012-2017) of medical students' SCT scores in the penultimate (Year 3) and final year (Year 4) of the undergraduate medical program. A repeated measure t-test indicated that the mean SCT scores for Year 4 students (M= 68.90, SD=7.79, N= 584) was higher than the mean SCT score for Year 3 students (M=64.03, SD=7.98, N=598). This difference in penultimate and final year students' mean SCT score, was statistically significant [$t_{(552)}=12.78$, $p< 0.001$]. A medium effect size was observed in the data, with Cohen's d repeated measures, pooled=0.544 (95%CI= 0.417 to 0.657). The means of SCT scores from 2012 to 2017 for the Year 3 and Year 4 students are represented in Figure 2.

Discussion

This study has provided evidence for the construct validity of SCT scores as a measure of clinical reasoning ability of undergraduate medical students. When the same set of SCT questions were given to senior medical students, junior GP registrars and experienced GPs, a significant upward progression of the SCT scores, from senior medical student level (relative novices) to practising GP clinician level (experienced clinicians), was noted. This suggests that GPs have more well-developed clinical reasoning skills, supporting the earlier observation that SCT scores tend to consistently increase with increasing level of training.[17] This result correlates with the study results showing progression of SCT scores from medical students to residency trainees in a Neurology training program.[1]

Although there was no statistical significance difference between the mean SCT scores for senior medical students and junior GP registrars, an upward trend was evident (from 64.90% to 68.36%). This could partially be explained by the

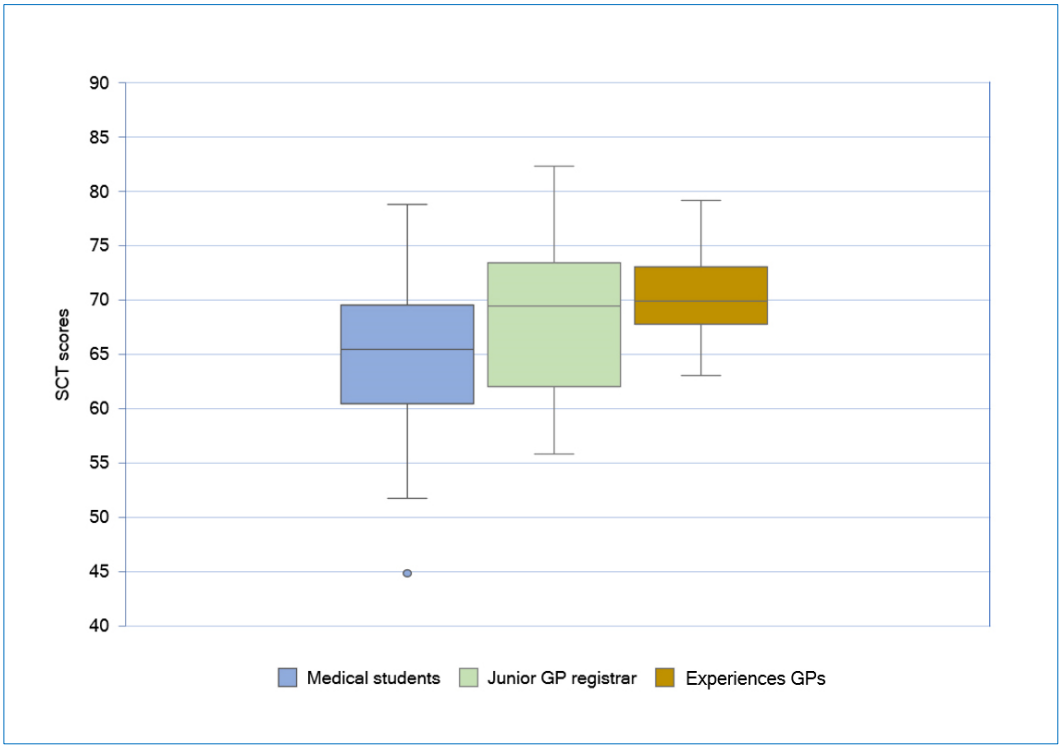


Figure 1. Comparison of SCT scores of medical students, junior registrars and practising GPs

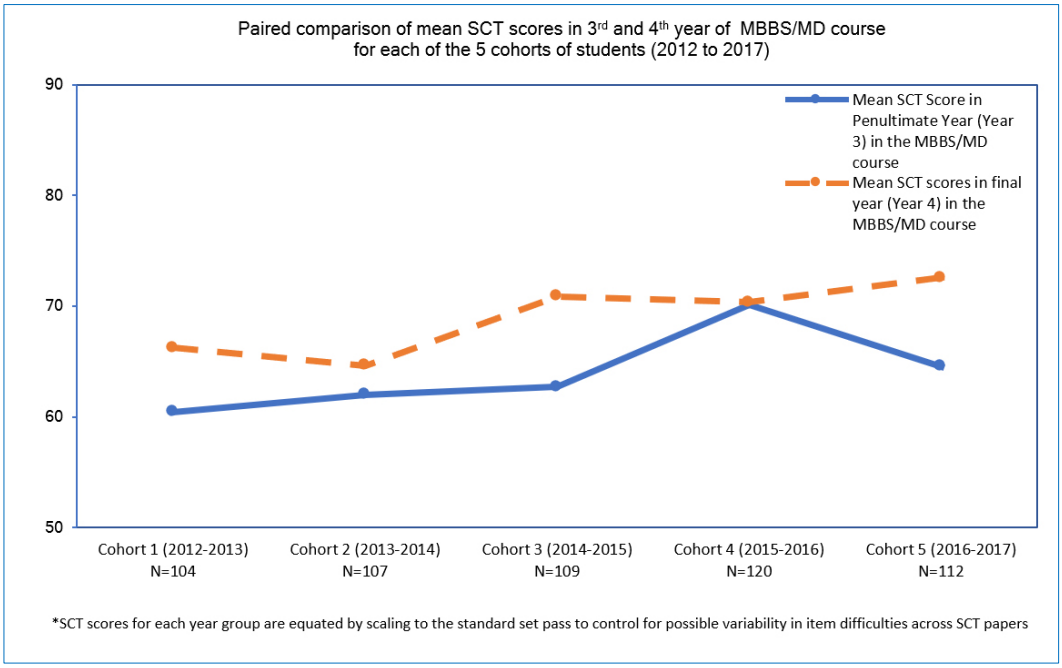


Figure 2. A paired comparison of mean SCT scores in Year 3 and Year 4 medical students for 5 cohorts (from 2012 to 2017)

fact that significant improvement in clinical reasoning with clinical experience is a progressive process occurs over a significant period of time. Subra et al. found that postgraduate students' clinical reasoning skills take time to develop and the largest improvement occurs during the first 18 months of training in general practice.[19] The smaller effect size could also be due to a plausible confounder – i.e. the medical students' more recent, and better experience (with more practice), with the assessment modality, compared to either the registrars or experienced GP participants.

The second part of the study using paired data from 2012 to 2017 demonstrated progression in medical students' SCT scores from Year 3 to Year 4. The effect size of 0.544 indicates that when medical students advance from Year 3 to Year 4, gaining more clinical experience, their mean SCT score also increases by 0.5 standard deviation.[27] A contrasting result

was reported from another Australian university with a 6-year undergraduate medical programme, where the SCT scores of Year 6 students in a formative assessment were compared to those of Year 5 students undertaking an end-of-year summative assessment.[6] In this instance, Year 6 students had less experience in answering SCT format questions. The significantly lower SCT means scores achieved by Year 6 students compared to those in Year 5 may highlight the benefit of prior experience with SCT items, and the potentially positive effect of sitting a high-stakes examination on candidate performance.

The results from the current study suggest that clinical experience does have an effect on performance in SCT, providing further support for the construct validity of scores from this format of assessment. Whilst previous studies have reported similar results at post-graduate medical education stage and shown progression of scores as trainees advance through their training in Neonatology and Psychiatry, there have been no reports of progression of SCT scores when the same set of SCT items are used to compare the clinical reasoning ability of medical students, junior doctors and experienced clinicians.[8,18] A study from Brazil has shown progression of SCT scores from students in the pre-clinical phase to those in the clinical phase (51.6% to 63.4%) using 10 clinical cases. However, the authors concluded that the implementation of this exam format is difficult in under-resourced institutions and have not followed up on these findings.[28]

Limitations of the study

In the summative assessment program of our medical school, Script Concordance Testing is a subset of the written paper, which limits each SCT section to 40 items only. The multiple-choice and short answer questions aim to test student knowledge, and ability to apply knowledge to clinical scenarios, whilst the SCT questions are included to test clinical reasoning. Including a greater number of SCT questions may help to elucidate whether there is a significant difference between medical student and junior GP registrar performance, as well as whether there is a significant difference between performance of junior GP registrars and experienced GP clinicians. It should also be noted that the SCT assessment is a high stakes examination for medical student participants, in contrast to registrar and GP participants in this study, where there is absolutely no stake in their participation in answering the SCT questions. Unequal sample size for each group used for comparison in the first part of this study should also be acknowledged as a potential limitation. This is particularly so for the sample size of junior registrar participants, which may, to a certain extent, explains the observation that while the scores of junior registrars were higher than senior medical students, collectively the difference in mean scores has failed to reach statistical significance. Nevertheless, the one-way ANOVA statistic used, is rather robust for comparisons involving unequal sample size in groups.[29] The findings are also limited in that the analysis was only performed on student results from one medical school.

More importantly, we acknowledge the fact that SCT scores are vulnerable to various validity threats and hence we are cautious not to over-claim with unrealistic inferences based on results from our limited data and simple convenient research design.[9,17,30,31] Nonetheless the current study adds to the limited available literature examining the progression of SCT scores with advancing clinical experience, especially in the undergraduate medical education setting.

A further study is underway to investigate the addition of a "think-aloud" written explanation to each SCT clinical scenario where the candidates are asked to explain their reasoning for choosing a particular response for each SCT question.[32,33] The response process validity of SCT scores as a measure of the clinical reasoning skills of undergraduate medical students would be enhanced if the majority of students chose the correct answer (to which the majority of experts agreed) for the correct reason, rather than providing correct answer-wrong reason responses. This qualitative data will add to the understandings of basis for any differences in SCT capability noted across the vertical continuum of medical education.

Further studies should look into the progression of clinical reasoning capabilities from Year 3 to Year 4 of the graduate-entry medical program in more than one medical school. A prospective longitudinal study involving a greater sample size of medical graduates would be more powerful in determining whether there is positive progression in clinical reasoning ability as measured by the SCT during junior doctor years, as compared with doctors doing fellowship training and subsequent clinical practice.

Conclusions

The increase in SCT scores of experienced GPs compared to medical students, and the higher SCT scores of final year medical students compared to their student peers in the penultimate year, support previous research findings that SCT scores consistently increase with increasing level of training. This study in one context of undergraduate medical education added further evidence to the body of literature concerning the construct validity of SCT as an assessment modality. Prospective longitudinal studies with larger sample sizes are recommended to further test the construct validity of SCT scores.

Acknowledgments

We would like to acknowledge Miss Eunice Lau for her support in collating the anonymous SCT examination data.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. The script concordance test: a new tool assessing clinical judgment in neurology. Can J Neurol Sci. 2009;36(3):326.
2. 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.
3. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: from theory to practice: AMEE Guide No. 75. Med Teach. 2013;35(3):184-93.
4. Deshpande S, Chahande J, Rathi A. Mobile learning app: a novel method to teach clinical decision making in prosthodontics. Educ Health (Abingdon). 2017;30(1):31.
5. Drolet P. Assessing clinical reasoning in anesthesiology: making the case for the script concordance test. Anaesth Crit Care Pain Med. 2015;34(1):5-7.
6. Duggan P, Charlin B. Summative assessment of 5th year medical students' clinical reasoning by script concordance test: requirements and challenges. BMC Med Educ. 2012;12(1):29.
7. Hamui M, Ferreira JP, Torrents M, Torres F, Ibarra M, Ossorio MF, et al. Script concordance test: first nationwide experience in pediatrics. Arch Argent Pediatr. 2018;116(1):e151-e155.
8. Kazour F, Richa S, Zoghbi M, El-Hage W, Haddad FG. Using the script concordance test to evaluate clinical reasoning skills in psychiatry. Acad 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. Am J Surg. 2012;203(4):530-4.
10. Phan SV. Cases in Psychiatry: a description of a multi-campus elective course for pharmacy students. Ment Health Clin. 2018;8(1):18-23.
11. Tsai TC, Chen DF, Lei SM. The ethics script concordance test in assessing ethical reasoning. Med Educ. 2012;46(5):527-.
12. Wan SH. Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine. Hong Kong Med J. 2015;21(5):455-61.
13. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7.
14. Carrière B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a script concordance test. Ann Emerg Med. 2009;53(5):647-52.
15. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52(4):281-302.
16. DeVon HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, et al. A psychometric toolbox for testing validity and reliability. J Nurs Scholarsh. 2007;39(2):155-64.
17. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance testing: a review of published validity evidence. Med Educ. 2011;45(4):329-38.
18. 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 Pediatric. 2016;16(6):e6.
19. Subra J, Chicoulaa B, Stillmunkes A, Mesthe P, Oustric S, Rouge Bugat ME. Reliability and validity of the script concordance test for postgraduate students of general practice. Eur J Gen Pract. 2017;23(1):209-14.
20. Cochran-Smith M, Donnell K. Practitioner inquiry: blurring the boundaries of research and practice. Handbook of complementary methods in education research. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 2006.
21. Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. BMC Med Inform Decis Mak. 2008;8(1):18.
22. Lubarsky S, Dory V, Meterissian S, Lambert C, Gagnon R. Examining the effects of gaming and guessing on script concordance test scores. Perspect Med Educ. 2018;7(3):174-81.
23. Wan MSH, Tor E, Hudson JN. Improving the validity of script concordance testing by optimising and balancing items. Med Educ. 2018;52(3):336-46.
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. Cohen J. A power primer. Psychol Bull. 1992;112(1):155-9.
26. Cohen J. Statistical power analysis for the behavioral sciences 2nd ed. New York: American Statistical Association; 1989.
27. Lenhard W, Lenhard A. Calculation of effect sizes. 2016 [cited 21 July 2019]; Available from: https://www.psychometrica.de/effect_size.html.
28. Roberti A, Roberti Mdo R, Pereira ER, Costa NM. Script concordance test in medical schools in Brazil: possibilities and limitations. Sao Paulo Med J. 2016;134(2):116-20.
29. Grace-Martin K. When unequal sample sizes are and are not a problem in ANOVA - the analysis factor. 2009 [Cited 21 July 2019]; Available from: https://www.theanalysisfactor.com/when-unequal-sample-sizes-are-and-are-not-a-problem-in-anova/.
30. 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? Int J Emerg Med. 2014;7:34.
31. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47(12):1175-83.
32. Power A, Lemay J-F, Cooke S. Justify Your Answer: The role of written think aloud in script concordance testing. Teach Learn Med. 2017;29(1):59-67.
33. Tedesco-Schneck M. Use of script concordance activity with the think-aloud approach to foster clinical reasoning in nursing students. Nurse Educ. 2019:44(5):275-277.

Synopsis of Chapter 5

The results reported in the published manuscript in this chapter provided further evidence of the construct validity of SCT scores, showing progressive improvement of scores from medical students (relative novice) from Year 3 to Year 4; and through to practising GPs (relative experts). The paper provided validity evidence and further support for the use of the SCT tool to assess clinical reasoning skills in medical education.

Chapter 6: Commentary: Expert responses in script concordance tests: A response process validity investigation

This chapter contains the invited commentary titled "Expert responses in script concordance tests: A response process validity investigation" published in Medical Education (ERA Journal) 2019;53(7):644-6.

Statement of Contribution by Others: please refer to Appendix 5

Foreword

The following article is an invited commentary in response to a paper published in *Medical Education* by Lineberry et al in 2019, investigating the response process of expert reference panels in SCT. (61) In the original paper, the authors raised concerns about the disagreements among experts in the panel and therefore the credibility or content accuracy of SCT items, and the instability of their responses in test-retest settings. This commentary has highlighted the positive trend of a 'deconstructed' approach to the study of the validation of SCT scores. It has also emphasised that all assessment tools, regardless of how well grounded they are on sound theoretical underpinning and empirical data, may demonstrate unintended issues on any given administration. Therefore, validity must be established repeatedly with adequate evidence collected from each administration, together with the deliberate exploration of what might be causing problems when unexpected findings arise.

Approval granted from Journal Editor for inclusion in the Thesis (Appendix 9).

Commentary: expert responses in script concordance tests: a response process validity investigation

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

There is substantial evidence that clinical decision making and medical problem solving by doctors depend to a large degree on probabilistic logic[1] and/or typicality of patient information with reference to doctors' activated illness scripts.[2] The Script Concordance Test (SCT) is a written assessment format designed specifically to assess individuals' performance on probability-related clinical information processing tasks. It presents candidates with a clinical scenario and requires them to consider a new piece of clinical information to determine the extent to which that information alters the probability of a particular diagnosis or appropriateness of a particular investigation or action.

The SCT is built upon sound conceptual and theoretical underpinnings.[3] A number of studies have explored the validity of SCT score interpretation, mostly comprising systematic gathering and documenting of evidence that SCT scores are indeed indicative of the soundness of candidates' clinical judgement.[4–6] The latest research using a 'deconstructed' approach to validation of SCT scores is a very positive trend that is helping shed some light on the many grey areas surrounding the validity of SCT scores.[6–8]

The SCT is built upon sound conceptual and theoretical underpinnings

The latest research using a 'deconstructed' approach to validation of SCT scores is a very positive trend that is helping shed some light on the many grey areas surrounding the validity of SCT scores

One such grey area derives from the fact that there is still limited study of response process validity (whether or not the responses of test takers suggest they share the same conception of the construct being measured as do the assessors). This is true for assessment generally, but is a particularly important issue for SCT designers, because the very rationale for SCT use is based on the assumption that candidates' answers reflect the cognitive operations involved in integrating newly presented patient information into existing medical knowledge structures to generate updated probabilities of a particular outcome.

Lineberry et al.'s study,[9] published in this issue, is an effort in this direction. The authors explored the response processes of experts to understand their divergent beliefs about how new clinical data alter the suitability of proposed actions and how they reacted to other experts' perspectives. Their study elicited varieties of expert responses other than those intended, providing evidence of construct irrelevant variance in the experts' response process.[10] These findings corroborate recent literature outlining plausible validity threats for SCT score interpretation.[11] In particular, empirical data from this study highlighted that typical SCT formats in which post-data belief changes by experts are interpreted without considering experts' pre-data belief run the risk of masking underlying agreement or disagreement between experts. Other significant findings reported include: (i) experts' disagreement with the proposed action in SCT items, raising concerns about the credibility or content accuracy of SCT items; and (ii) instability of experts' responses, indicating a threat to the test–retest reliability of SCT scores. The authors discuss the challenge of balancing the tension between maintaining authenticity in reflecting 'uncertainty' in clinical decision making and, at the same time, ensuring content accuracy in SCT items.

In collecting these data, Lineberry et al. acknowledge their SCT cases were adapted from real patient histories, with rich details and findings. As such, they may have diverged from the usual SCT guidance to be 'brief' and 'ill-defined'.[3,12,13] This could be a

[1]School of Medicine, University of Notre Dame, Sydney, New South Wales, Australia
[2]Faculty of Health and Medical Sciences, University of Adelaide, Adelaide, South Australia, Australia

Correspondence: Siu Hong Wan, Medical Education Unit, The University of Notre Dame, 160 Oxford Street, Darlinghurst, New South Wales 2010, Australia. Tel: 0282044479;
E-mail: michael.wan@nd.edu.au

doi: 10.1111/medu.13889

critically important feature determining the extent to which SCTs are implemented in a way that yields valid scores. In our experience constructing SCT items, using simple and ill-defined case scenarios to test core concepts in clinical reasoning in medicine, only 20–30% of SCT items have generally been discarded or modified because of discordance in response pattern amongst clinicians (i.e. extreme inconsistencies among experts).[14]

This issue is not raised in an effort to facilely dismiss Lineberry et al.'s results, given that is purely a speculative hypothesis at the moment. Rather, it is mentioned because Lineberry et al.'s findings remind us of the broader issue that it is important to remain aware of the fact that all assessment tools, regardless of how well grounded they are on sound theoretical underpinning and empirical data, may demonstrate unintended issues on any given administration. Validity must be established repeatedly with adequate evidence collected on each administration and deliberate exploration of what might be causing problems when unexpected findings arise. Evidence supporting the use of test scores should be documented over time, from multiple sources, consistent with the contemporary conception of validity as a unitary construct.[15–17] That is, validation should be an ongoing process forming part of the fabric of all assessment initiatives,[18] but particularly in the context of high-stakes summative assessments of learning.

All assessment tools, regardless of how well grounded they are on sound theoretical underpinning and empirical data, may demonstrate unintended issues

At the structural level, recent calls for a move towards a programmatic perspective on assessing competence is a paradigm shift in the right direction towards a more sustainable and constructive landscape in medical education. This more continuous form of assessment makes it all the more imperative that we adopt a continuous form of validation practices. The rich information that SCTs can provide, as discussed by Lineberry et al., can be optimised for learning and be meaningfully aggregated to inform progress decisions for trainees, but only if care is put into ensuring that the scores reflect what the theory intends.[19–21] The post-scoring debrief and debate by the expert reference panel, used by Lineberry et al. in this paper, provides an excellent example of a counter-measure to be used against validity threats that could simultaneously serve as a useful continuing professional development activity for clinicians, test developers and educators alike. Engaging in such activity may turn controversial SCT cases into valuable stimuli for learning, hence achieving and role modelling the goal of authentically reflecting the complexity of medical decision making.[21,22]

With this shift towards programmatic assessment, the rich information that SCTs can provide, as discussed by the authors, can be optimised for learning.

The post-scoring debrief and debate by the expert reference panel may turn controversial SCT cases into valuable stimuli for learning

In sum, although Lineberry et al.'s findings might be considered a negative mark on the validity evidence for SCT use, we argue that the authors' research approach more constructively provides a strategy for enabling a longitudinal exploration of the development of clinical reasoning in both learners and experts. Panels comprised of undergraduate or postgraduate learners can learn from reflection on proposed actions (pre and post) as well as the responses of peers and experts. This could provide valuable understanding of what stimulates clinical reasoning ability in learners during their professional development and what maintains or furthers clinical reasoning ability in those who are more well established.

REFERENCES

1 Kreiter CD. A Bayesian perspective on constructing a written assessment of probabilistic clinical reasoning in experienced clinicians: assessing clinical reasoning. *J Eval Clin Pract* 2017;**23** (1):44–8.
2 Lubarsky S, Gagnon R, Charlin B. Script concordance test item response process: the argument for probability versus typicality. *Adv Health Sci Educ* 2012;**17** (1):11–3.
3 Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The script concordance test: a tool to assess the reflective clinician. *Teach Learn Med* 2000;**12** (4):189–95.
4 Lubarsky S, Vleuten CPMVD, Charlin B, Chalk C, Cook DA. Script concordance testing: a review of published validity evidence. *Med Educ* 2011;**45** (4):329–38.
5 Subra J, Chicoulaa B, Stillmunkes A, Mesthe P, Oustric S, Bugat MER. Reliability and validity of the script concordance test for postgraduate students of general practice. *Eur J Gen Pract* 2017;**23** (1):209–14.

6 Wan MS, Tor E, Hudson JN. Improving the validity of script concordance testing by optimising and balancing items. *Med Educ* 2018;**52** (3):336–46.
7 Lubarsky S, Dory V, Meterissian S, Lambert C, Gagnon R. Examining the effects of gaming and guessing on script concordance test scores. *Perspect Med Educ* 2018;**7** (3):174–81.
8 Wan SH, Duggan P, Tor E, Hudson JN. Association between candidate total scores and response pattern in script concordance testing of medical students. *Focus Health Prof Educ* 2017;**18** (2):26–35.
9 Lineberry M, Hornos E, Pleguezuelos E, Mella J, Brailovsky C, Bordage G. Experts' responses in script concordance tests: a response process validity investigation. *Med Educ* 2019; https://doi.org/10.1111/medu.13814.
10 Whitely SE. Construct validity: construct representation versus nomothetic span. *Psychol Bull* 1983;**93** (1):179–97.
11 Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. *Med Educ* 2013;**47** (12):1175–83.
12 Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. *Med Educ* 2012;**46** (6):552–63.
13 Duggan P, Charlin B. Summative assessment of 5th year medical students' clinical reasoning by script concordance test: requirements and challenges. *BMC Med Educ* 2012;**12** (1):29.
14 Wan SH. Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine. *Hong Kong Med J* 2015;**21** (5):455–61.
15 Crooks TJ, Kane MT, Cohen AS. Threats to the valid use of assessments. *Assess Educ* 1996;**3** (3):265–86.
16 Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas* 2013;**50** (1):1–73.
17 Messick S. Meaning and values in test validation: the science and ethics of assessment. *Educ Res* 1989;**18** (2):5–11.
18 Tor E, Steketee C, Mak D. Clinical audit project in undergraduate medical education curriculum: an assessment validation study. *Int J Med Educ* 2016;**7**:309–19.
19 Heeneman S, Pool AO, Schuwirth LWT, van der Vleuten CPM, Essen EW. The impact of programmatic assessment on student learning: theory versus practice. *Med Educ* 2015;**49** (5):487–98.
20 Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ* 2012;**46** (1):38–48.
21 Van Der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39** (3):309–17.
22 Power A, Lemay J-F, Cooke S. Justify your answer: the role of written think aloud in script concordance testing. *Teach Learn Med* 2017;**29** (1):59–67.

Synopsis of Chapter 6

This invited commentary on published research findings pertaining to the impact of the expert reference panel's response process on the validity of SCT scores, emphasised the importance of continual validation of SCT scores when they are used as measures of the clinical reasoning ability of medical trainees. (61) The commentary highlighted the need for analysis of the underlying thought processes and clinical reasoning, of both candidates and experts in the reference panel, in deriving their clinical decision in response to each scenario in the SCT items. It also discussed the move towards the era of programmatic assessment, where multiple assessment points are used to allow a longitudinal assessment of a student accompanied by rich feedback, post-scoring debriefs and debate by the experts in the panel. The latter process may turn controversial SCT cases into valuable stimuli *'for'* learning. This forms part of the continuum of consequences evidence for the construct validity of SCT scores.

Chapter 7: Examining response process validity of Script Concordance Testing: a think-aloud approach

This chapter contains the paper titled "Examining response process validity of Script Concordance Testing: a think-aloud approach" published in IJME (ERA Journal) 2020;11:127-135.

Statement of Contribution by Others: please refer to Appendix 6

Foreword

In response to the plausible threats to the validity of the SCT scores reported in the literature, specifically in students' response process, the following chapter reports and discusses a preliminary investigation into the use of a new written 'think-aloud' approach to explore medical students' response process in answering SCT questions. Students' response process validity has been quantified as *true positive* rates – the percentage of full and partial credit responses which were derived based on correct clinical reasoning; and the *true negative* rate – the percentage of responses with no credit that were derived based on incorrect reasoning. Students' perceptions of the usefulness of this approach in enhancing their learning of clinical reasoning skills, was also briefly explored.

International Journal of Medical Education. 2020;11:127-135
ISSN: 2042-6372
DOI: 10.5116/ijme.5eb6.7be2

Examining response process validity of script concordance testing: a think-aloud approach

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

[1]School of Medicine, The University of Notre Dame Australia, Australia
[2]Faculty of Health and Medical Sciences, University of Adelaide, Australia

Correspondence: Michael Wan, 160 Oxford Street, Darlinghurst, NSW 2010, Australia. Email: michael.wan@nd.edu.au

Accepted: May 09, 2020

Abstract

Objectives: This study investigated whether medical student responses to Script Concordance Testing (SCT) items represent valid clinical reasoning. Using a think-aloud approach students provided written explanations of the reasoning that underpinned their responses, and these were reviewed for concordance with an expert reference panel.

Methods: A set of 12, 11 and 15 SCT items were administered online to Year 3 (2018), Year 4 (2018) and Year 3 (2019) medical students respectively. Students' free-text descriptions of the reasoning supporting each item response were analysed, and compared with those of the expert panel. Response process validity was quantified as the rate of true positives (percentage of full and partial credit responses derived through correct clinical reasoning); and true negatives (percentage of responses with no credit derived through faulty clinical reasoning).

Results: Two hundred and nine students completed the online tests (response rate = 68.3%). The majority of students who had chosen the response which attracted full or partial credit also provided justifications which were concordant with the experts (true positive rate of 99.6% for full credit; 99.4% for partial credit responses). Most responses that attracted no credit were based on faulty clinical reasoning (true negative of 99.0%).

Conclusions: The findings provide support for the response process validity of SCT scores in the setting of undergraduate medicine. The additional written think-aloud component, to assess clinical reasoning, provided useful information to inform student learning. However, SCT scores should be validated on each testing occasion, and in other contexts.

Keywords: Script concordance testing, response process validity, written think-aloud, assessment, clinical reasoning

Introduction

Clinical reasoning is a cognitive process where the clinician collects information from the history, physical examination and/or investigations related to a patient presentation to come to a conclusion about the patient's health situation. Thereby this process allows for the implementation of appropriate intervention and management.[1] As diagnostic errors are often related to problems with clinical reasoning, the Accreditation Council for Graduate Medical Education (ACGME) has recently urged educators to include clinical reasoning as a core competency in undergraduate and graduate medical education.[2] This was in response to relatively slow progress in improving the teaching and assessment of clinical reasoning, including the development of specific assessment tools, as well as research and innovation in clinical reasoning education.[2] One such explicit tool, Script Concordance Testing (SCT), was described by Charlin in 2000.[3] It has since been used to assess clinical reasoning in health professional education.[4,5] In Medicine, SCT has been implemented in Paediatrics, Neurology, Emergency Medicine and Psychiatry disciplines.[6-11] A body of research informing the use of SCT to assess clinical reasoning has gradually developed over the past decade or so.

In each SCT, an authentic but non-complex clinical scenario/vignette is presented and students are asked to assess whether an additional piece of information increases or decreases the probability/appropriateness of the diagnosis, investigation or management in the context of uncertainty.[12] The additional information could be in the form of further history symptoms, physical examination signs, or investigations or imaging findings.[13] The response to each item is recorded on a 5-point scale from '-2': much less likely/appropriate; '-1': less likely/appropriate; '0': neither less or more, likely/appropriate; '+1': more likely/appropriate; '+2': much more likely/appropriate.[12,13] For example, a 45-year-old man presenting to the Emergency Department with acute onset of chest pain and shortness of breath for 5 hours. The student is asked to determine whether the additional finding of 'unilateral swelling with dilated veins in the right leg' would make

the diagnosis of 'pulmonary embolism' much less or less likely, neither less or more likely, more or much more likely. In order to answer the item, the student will need to activate 'illness scripts' in his or her mind, which have been constructed based on previous clinical encounters/experiences.[14] In scoring these SCT items, student responses are compared to responses from a panel of experts using the same 5-point scale. The classical aggregated (weighted) scoring system is used.[13] If a student's response 'concords' with that of the majority of the expert panel (i.e. the modal response of the expert reference panel), a score of '1' (full credit) is awarded reflecting that the consensus reasoning has been applied. A partially weighted credit is awarded if the student's response 'concords' with a minority of the panel, reflecting a difference in interpretation that may still be clinically valuable and worthy of partial credit. Finally, no (0) credit is awarded if none of the experts have chosen the particular response (Table 1).[13] It is the partially weighted credit in SCT scoring that differentiates it from scoring of classical multiple-choice questions, where only one single best answer response will attract the full one mark. This unique scoring system in SCT acknowledges important real-world clinical situations, where clinicians often interpret data and make alternate clinical decisions, especially under uncertain conditions.

Table 1. The formula to calculate the aggregated (weighted) scores for each SCT item

Score Key	-2	-1	0	+1	+2
Number of experts in the panel choosing the response (out of 10)	0	0	1	2	7
Formula	0/7	0/7	1/7	2/7	7/7
Student score	0	0	0.14	0.29	1

While SCT has been shown to be a valid and reliable assessment tool in various examination settings,[15-18] recent research has questioned the plausible threats to the validity of the SCT scores, specifically in relation to student and/or expert reference panel member response processing. The classical SCT item format only captures the student's response along the 5-point scale. The actual clinical reasoning and thought processing involved in choosing a particular response is not recorded or examined. To explore student's response processes in the assessment of clinical reasoning, an approach from education, the think-aloud approach, has been applied to health professional education. This approach has proved useful to allow medical, pharmacy and nursing students' thought processes to be examined.[15-26] Trainees were asked to either write down or verbalise their thought processes in relation to decision-making when choosing the answer. Pinnock and colleagues found this approach useful in helping both medical students and supervisors learn and teach clinical reasoning in the clinic environment.[24] In the critical care setting, the think-aloud protocol had been used during ICU rounds to identify strengths and weaknesses concerning the trainees' clinical decision-making processes.[25] The think-aloud method has also been used to improve the training of community pharmacists when reviewing medications for patient safety.[19] Johnsen and colleagues noted that the verbal think-aloud approach could help to understand nurses' clinical reasoning in real-life clinical practice and hence provided nurse educators with ways to improve teaching methods in Nursing.[21] Another study in a paediatric nursing course found that the written think-aloud approach (followed by small group discussions) could foster the learning of clinical reasoning. Students reported increased confidence, as well as valuing the importance of in-depth discussion associated with the items.[27]

Recently, the think-aloud approach has been used to further elucidate the utility of SCT in assessing the clinical reasoning of medical undergraduates and postgraduate trainees.[27,28] The process of asking students to justify their reasons for choosing a particular SCT response option, was in response to Kreiter's critique 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.[29] Power and colleagues recently used the think-aloud approach to understand the actual response process of paediatric postgraduate trainees in six SCT cases that covered diagnosis, investigation and treatment. They concluded that the written think-aloud approach could identify incorrect clinical reasoning with correct SCT responses, sound clinical reasoning with sub-optimal SCT responses and misinterpretation of the SCT question.[28] Their study suggested that the think-aloud approach could strengthen the quantitative assessment method provided by the classical SCT. It was valued as an approach providing assessment *for*, as well as *of*, trainees' learning. Indeed, in response to this, Lubarsky and colleagues 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.[30]

As mentioned above, in SCT, as in all other assessment items in multiple-choice response format, the actual reasoning behind the selection of a particular response option by individual examinees is never clear. The validity of score interpretation is based on the assumption that correct responses by examinees were derived based on appropriate and correct reasoning processes.[31] Use of the written think-aloud approach offered the potential to explore whether a student's response to each SCT item is underpinned by correct reasoning, consistent with that of the expert reference panel. Thus we applied a similar approach to specifically explore the 'response process' validity of SCT scores in assessing the clinical reasoning of senior medical students.[32] Unlike Power and colleagues' study,[28] our study investigated the response process validity of the written think-aloud approach in the under-graduate medical student setting with SCT questions across multiple disciplines.

This study aimed to investigate the 'response process' validity of SCT scores in assessing the clinical reasoning of

senior medical students. Students were asked to explain in a text box, the thought process involved in deriving the particular response option selected for each SCT item. The study sought to answer two questions: 1) Are full and partial credit responses from students derived through correct clinical reasoning; and 2) Are responses with no credit indeed a result of faulty clinical reasoning?

Methods

Study design

In this descriptive study, a set of 12, 11 and 15 SCT items were administered online to Year 3 (2018), Year 4 (2018) and Year 3 (2019) students, respectively. This was an online test offered by the school's assessment team to prepare students for the year-end summative SCT examination (2018-2019). The SCT items were selected from an item bank of 500 items. The content of each item was mapped to the curriculum of the two clinical years covering Medicine, Surgery, Paediatrics, Psychiatry, Women's Health and General Practice disciplines. Each SCT scenario had been reviewed by the relevant discipline leads and the assessment academics to ensure content validity. The expert panel used for scoring the items consisted of specialists in the relevant disciplines and general practitioners who were directly involved in the teaching of the students. In constructing the final sets of SCT items, any item with inconsistent panelist responses (bi-modal or uniform divergence responses) was modified or discarded to optimise the test before implementing the online test. This step aimed to improve the validity of the assessment tool.[12]

Detailed descriptions of the format of SCT and the scoring system were given to the students at the beginning of the test. To help the student to understand and improve their clinical reasoning and decision-making skills, each student was asked to record online, the reasons behind each of their chosen answers for the SCT items (written think-aloud approach) before choosing the answers according to the 5-point scale.

Participants

All students in Year 3 and Year 4 from 2018, and in Year 3 in 2019, of the medical program, were invited to participate voluntarily via an announcement on the university's student learning portal. Participant Information was presented online, and consent was obtained by the students clicking the "agree to include the anonymised data for analysis in the study" key. This approach allowed the students to continue to attempt practice SCT items and receive the usual feedback even if their responses were not being collected for this study.

Data collection

In this study, the online practice tests were delivered via a free online survey tool. The students' answers were compared with those of the expert panel members (n=15). The classical SCT weighted aggregate scoring method was used for scoring. A full credit was given if a student's response was the same as the expert panel's modal response, and a partial credit was given if a student's response concorded with the minority of the panel according to the formula as shown in Table 1 above. The free text explanation of the clinical reasoning behind choosing each answer was also collected. The keyed responses and explanations data were transferred to a spreadsheet for coding and anonymous analysis. Immediately following the test, students were provided (online) with the responses which attract full and partial credit and the experts' clinical reasoning behind each decision. This was followed by a separate face-to-face feedback session where significantly incorrect clinical reasoning or misinterpretation related to the students' written think-aloud free text entries were explained and discussed with the cohort. This session aimed to improve student clinical reasoning skills and help them to better prepare for the summative examination.

Ethics approval for the study was obtained from the University's Human Research and Ethics Committee (#019023S). All participant data and free text explanations were collected anonymously via the online survey tool.

Data analysis

Response process validity was quantified as the true positive (TP) rate, i.e. percentage of full and partial credit responses derived through correct clinical reasoning; and true negative (TN) rate, i.e. percentage responses with no credit derived through incorrect/faulty clinical reasoning.

The first author analysed students' free-text justifications for their answers for each of the SCT items. For each SCT item, student's clinical reasoning explanation was compared with the experts' consensus reasoning, to evaluate the extent of concordance between the two, i.e. students and expert clinicians from the reference panel. Students' written think-aloud explanations were coded into six categories: A) Full credit response derived based on correct reasoning, concordant with the experts' reasoning (true positive in full credit responses); B) Partial credit response derived based on correct reasoning, concordant with the experts' reasoning (true positive in partial credit responses); C) Full or partial credit response derived based on incorrect/faulty reasoning as compared with the experts (false positive in both full and partial credit responses); D) Response that received no credit through faulty clinical reasoning (true negative); E) Response that received no credit but free text justification indicates correct reasoning concordant with the experts' reasoning, due to mis-selection of the score keys (false negative); F) Response that received no credit even though free text justification indicates correct reasoning, because none of the expert reference panel members had selected that particular response option (false negative). According to the above categories, the percentage of true positives and true negatives were calculated for the student responses analysed.

Results

Students' response process

The participation rate was 68.3% (N = 209). A total of 38 SCT items (12 for Year 3 in 2018; 11 for Year 4 in 2018 and 15 for Year 3 in 2019), with 2,695 student responses were analysed. Of all the 1,679 responses provided by students to each of the SCT items which attracted a full credit (based on the extent of concordance with the expert panel's responses, i.e. the modal response), 1,673 were based on correct clinical reasoning (Category A – True Positives in full credit responses). Of the 700 responses which attracted a partial credit, 696 were based on correct clinical reasoning concordant with the experts (Category B – True Positives in partial credit responses). Ten responses which were awarded full or partial credits were derived based on incorrect/faulty clinical reasoning (Category C – False Positives). Of the 315 responses which attracted no credit, 312 were based on incorrect clinical reasoning (Category D – True Negatives). Two student participants (both in the Year 3 cohorts) had chosen the wrong response option despite correct clinical reasoning due to mis-selection of the wrong answer key (Category E – False Negatives). Three responses had the correct clinical reasoning but received no credit because none of the experts had selected that particular response option (Category F – False Negatives).

As mentioned above, the majority of students who had chosen the answer which attracts full or partial credit also provided justifications which were concordant with the experts (true positive rate of 99.6% for full credit and 99.4% for partial credit answers respectively). The majority of answers that attracted no credit were based on incorrect clinical reasoning (true negative rate of 99.0%).

Examples of students' free-text explanations (direct quotes) of their clinical reasoning with respect to the full or partial credit responses in each Category (A to F) are represented in Appendix 1.

Other findings

Reviewing the written think-aloud responses as part of the SCT test optimisation process allowed the experts/academics to discuss and modify any items that were flawed or prone to misinterpretation. The following example demonstrates how an SCT item on the investigation was modified after reviewing the written think-aloud explanation by students. The clinical scenario was a 45-year-old man who presents to the Emergency Department with a 3-day history of epigastric pain. The question asked whether the finding of the fact that the pain could be relieved by antacids would make ordering endoscopic examination less or more appropriate. The expert panel's modal answer was 'much less appropriate' (-2) as the procedure is invasive and would only be indicated if the patient had symptoms of anaemia, weight loss or poor response to medical treatment with antacid or proton pump inhibitors. However, the analysis of free-text responses revealed a significant number of students assumed that the patient has recurrence or persistence of epigastric pain; and therefore selected 'more appropriate' (+1) or 'much more appropriate' (+2) where no credits were awarded (Category F). For re-administration of the SCT question, the first presentation of the symptoms was clarified. The clinical scenario was modified to read 'a 45-year-old man presents to the General Practice with a 3-day history of epigastric pain. He has no previous history of similar pain'.

Discussion

This study sought to explore the response process validity of SCT scores as a proxy measure for the clinical reasoning ability of senior medical students through the written think-aloud approach. Most students seemed to have applied correct clinical reasoning in deriving responses which attract credit in the SCT test. The rate of true positives was 99.6% in full credit responses and 99.4% in partial credit responses. The true negative rate was 99.0%, whereby the students' responses based on faulty clinical reasoning did not earn any credit under the aggregated partial credit scoring model. There is currently no other study in the SCT literature on examinees' response process validity which quantifies the results as the rate of true positives and true negatives.

The fact that a few student responses (6 of 1,679 = 0.4%) had attracted a full credit despite incorrect/faulty clinical reasoning (false positives) suggested there was a potential response process validity threat to SCT scores interpretation due to a construct irrelevant variable. However, the rate of false positives was low. The example in Category C, as presented in the Appendix section, demonstrated that the SCT response format could possibly have a masked misconception by the student, despite the concordance with the expert panel response. Addressing misconceptions such as these in face-to-face discussions, after a written think-aloud approach, can provide students with powerful and timely learning of clinical reasoning. This can also be useful for educators and the expert reference panel to improve questions with ambiguity to avoid confusion and misinterpretation by students.

Recent research highlights that, for more complex and controversial clinical scenarios, the expert panel's modal responses could be variable and even inconsistent over time. Lineberry and colleagues[33] reported threats to response process validity due to variable expert panel consensus, but this occurred with complex and controversial cases in the postgraduate setting. Variability in the expert panel consensus is less likely with the use of simple classical SCT scenarios/cases.[34] The latter was used in the current study, and likely explain the very low rate of student discordance with panel responses. Care should also be taken in selecting SCT scenarios, to ensure they introduce sufficient level of 'uncertainty' to fit the conceptual underpinning of SCT (rather than a definitive answer). As described in the Results section, by reviewing the think-aloud responses of the students as part of the SCT test optimisation process, any items that are flawed or confusing can be modified for future administration.

Interestingly, on a few occasions, the students in Year 3 chose the wrong response option despite providing the correct underlying clinical reasoning and interpretation of the item (Category E). This could be due to unfamiliarity with the 5-point response scale of SCT, which could lead to confusion in choosing between the keys of '-2' and '+2' or '-1' and '+1'. More practice in answering SCT items could have minimised this, as this effect was not apparent for the Year 4 test. This is likely due to the fact that Year 4 students had previously been exposed to the SCT format.

Think-aloud is a very useful approach for SCT validation research, particularly in gathering evidence for response process validity in this multiple-choice assessment format. Think-aloud is also a powerful add-on mechanism to improve the educational impact of SCT as one assessment modality in the programmatic assessment. Feedback from the approach can support learners and facilitate further learning. As Power and colleagues demonstrated, in a formative assessment setting, students have the opportunity to better understand the underlying correct clinical reasoning through debriefing/feedback sessions conducted by their teachers.[28] The rich information potentially provided by SCT can be optimised for learning if care is put into ensuring that the scores reflect what the theory intends. The think-aloud approach and post-scoring debrief offered to students in the current study, provided an example of a counter-measure against validity threats and a stimulus for learning.[34]

A simple short post-test evaluation survey (unpublished) revealed that many participants found that the think-aloud approach with the expert panel's clinical reasoning feedback was helpful for supporting their learning by comparing their answers with the experts. The following anonymous quote from one student illustrated student perception that the debriefing, in explaining the expert panel's reasoning for each SCT item, was useful: by writing the explanation of why the investigation is appropriate and then comparing my thought process with the experts was invaluable for my learning in clinical decision making (Year 4 student).

From a programmatic assessment perspective, if documented systematically and aggregated meaningfully, the rich information from the written think-aloud in SCT can also inform important decision-making for student/trainee progression in training programs.[34] In high stakes summative examinations using SCT, marking the think-aloud components of the answers (although requiring additional resources for manual marking) may provide additional information in relation to student understanding of a given scenario. In addition to supporting the response process validity of SCT for assessment of medical undergraduates, the approach can facilitate student learning of clinical reasoning.

Limitations

The study was conducted in one medical school with two years of data (2018-2019) only, and there were limited numbers of SCT items in each test administered. However, score reliability was not as critical in this study, as it aimed to investigate the clinical reasoning underpinning student responses to SCT items, rather than students' overall aggregate scores in each SCT assessment (for pass-fail decisions). The online formative test items were reused from previous years, and therefore, some students may have been exposed to these items if they had been passed on by their senior peers. However, as this was a formative practice opportunity for students and was anonymous, the likelihood of deliberately preparing for such an examination or using an open-book approach was unlikely. In a voluntary setting with a 68.3% participation rate, lower-performing students might be under-represented. However, the very similar average SCT score between the formative test and subsequent summative examination (69% vs 67% respectively) could support the fact that the sampling of the cohort in the current study was representative. The formative nature of this study might limit the interpretation of the results, but to extend such written think-aloud answers in the summative setting without any actual scoring impact on the free text explanation would have been unfair to the students.

This study investigated a phenomenon in its natural setting, i.e. the cognitive process which underpins individual examinees' responses to each SCT item. Data gathered from this study facilitated a better understanding of the written think-aloud approach to answering SCT items in the undergraduate medical program setting, adding to research into clinical reasoning education.

Future directions

Collaboration with national and international institutions in further research of the think-aloud approach in answering SCT would provide more insight into the response process validity. Further studies using student focus groups could explore students' underlying thought process and thinking, in choosing between the various response-keys on the 5-point scale to ensure the responses are used correctly. There is increasing interest in the use of SCT in medical ethics, and the addition of the think-aloud process to student responses to SCT ethics items would be valuable for later group discussion of ethical dilemmas.[35]

Conclusions

Although a plausible response process validity threat to SCT score interpretation could arise due to a construct irrelevant variable, this study using a written think-aloud approach in a formative SCT setting in one medical school, demonstrated that the likelihood was relatively low. The finding that the majority of the student keyed-responses corresponded to the correct think-aloud clinical reasoning in various clinical disciplines added further evidence to support the response process validity of SCT scores. The findings have demonstrated that the use of SCT with an additional written think-aloud approach can be a very useful assessment modality for providing rich information to guide further learning.

The study has supported the use of SCT as an explicit tool to assess clinical reasoning in undergraduate medical education. However, validation of SCT scores requires ongoing effort. They should be validated on each testing occasion, and in other contexts.

Acknowledgements

We would like to acknowledge Miss Eunice Lau for her support in collating the anonymous SCT examination data and Dr Cassy Richmond for her editing input.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Krairiksh M, Anthony MK. Benefits and outcomes of staff nurses' participation in decision making. J Nurs Adm. 2001;31(1):16-23.
2. Connor DM, Durning SJ, Rencic JJ. Clinical reasoning as a core competency. Acad Med. 2019; Online ahead of print.
3. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The script concordance test: a tool to assess the reflective clinician. Teach Learn Med. 2000;12(4):189-95.
4. Dawson T, Comer L, Kossick MA, Neubrander J. Can script concordance testing be used in nursing education to accurately assess clinical reasoning skills? J Nurs Educ. 2014;53(5):281-6.
5. 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.
6. Carrière B. Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a script concordance test. Ann Emerg Med. 2009;53(5):647-52.
7. Claessens YE, Wannepain S, Gestin S, Magdelein X, Ferretti E, Guilly M, et al. How emergency physicians use biomarkers: insights from a qualitative assessment of script concordance tests. Emerg Med J. 2014;31(3):238-41.
8. Hamui M, Ferreira JP, Torrents M, Torres F, Ibarra M, et al. Script concordance test: first nationwide experience in pediatrics. Arch Argent Pediatr. 2018;116(1):E151-E5.
9. Kazour F, Richa S, Zoghbi M, El-Hage W, Haddad FG. Using the script concordance test to evaluate clinical reasoning skills in psychiatry. Acad Psychiatry. 2017;41(1):86-90.
10. Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. The script concordance test: a new tool assessing clinical judgement in neurology. Can J Neurol Sci. 2009;36(3):326.
11. Talvard M, Olives JP, Mas E. Assessment of medical students using a script concordance test at the end of their internship in pediatric gastroenterology. Arch Pediatr. 2014;21(4):372-6.
12. Wan M. Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine. Hong Kong Med J. 2015;21(5).
13. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: from theory to practice: AMEE guide no. 75. Med Teach. 2013;35(3):184-93.
14. Charlin B, Brailovsky C, Leduc C, Blouin D. The diagnosis script questionnaire: a new tool to assess a specific dimension of clinical competence. Adv Health Sci Educ Theory Pract. 1998;3(1):51-8.
15. Gagnon R, Charlin B, Lambert C, Carrière B, Van Der Vleuten C. Script concordance testing: more cases or more questions? Adv Health Sci Educ Theory Pract. 2009;14(3):367-75.
16. Humbert AJ, Johnson MT, Miech E, Friedberg F, Grackin JA, Seidman PA. Assessment of clinical reasoning: a script concordance test designed for pre-clinical medical students. Med Teach. 2011;33(6):472-7.
17. 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. Am J Surg. 2012;203(4):530-4.
18. Wan MS, Tor E, Hudson JN. Improving the validity of script concordance testing by optimising and balancing items. Med Educ. 2018;52(3):336-46.
19. Croft H, Gilligan C, Rasiah R, Levett-Jones T, Schneider J. Thinking in pharmacy practice: a study of community pharmacists' clinical reasoning in medication supply using the think-aloud method. Pharmacy. 2017;6(1):1.
20. Forsberg E, Ziegert K, Hult H, Fors U. Clinical reasoning in nursing, a think-aloud study using virtual patients – a base for an innovative assessment. Nurse Educ Today. 2014;34(4):538-42.
21. Johnsen HM, Slettebø Å, Fossum M. Registered nurses' clinical reasoning in home healthcare clinical practice: a think-aloud study with protocol analysis. Nurse Educ Today. 2016;40:95-100.
22. Lee J, Lee YJ, Bae J, Seo M. Registered nurses' clinical reasoning skills and reasoning process: a think-aloud study. Nurse Educ Today. 2016;46:75-80.
23. McAllister M, Billett S, Moyle W, Zimmer-Gembeck M. Use of a think-aloud procedure to explore the relationship between clinical reasoning and solution-focused training in self-harm for emergency nurses. J Psychiatr Ment Health Nurs. 2009;16(2):121-8.
24. Pinnock R, Fisher TL, Astley J. Think aloud to learn and assess clinical reasoning. Med Educ. 2016;50(5):585-6.
25. Siddiqui S. 'Think-aloud' protocol for ICU rounds: an assessment of information assimilation and rational thinking among trainees. Med Educ Online. 2014;19(1):25783.
26. Verkuyl M, Hughes M, Fyfe MC. Using think aloud in health assessment: a mixed-methods study. J Nurs Educ. 2018;57(11):684-6.
27. Tedesco-Schneck M. Use of script concordance activity with the think-aloud approach to foster clinical reasoning in nursing students. Nurse Educ. 2018:1.
28. Power A, Lemay J-F, Cooke S. Justify your answer: the role of written think aloud in script concordance testing. Teach Learn Med. 2017;29(1):59-67.
29. Kreiter CD. Commentary: The response process validity of a script concordance test item. Adv Health Sci Educ Theory Pract. 2012;17(1):7-9.
30. Lubarsky S, Gagnon R, Charlin B. Script concordance test item response process: the argument for probability versus typicality. Adv Health Sci Educ Theory Pract. 2012;17(1):11-3.
31. Kane MT. Validating the interpretations and uses of test scores. Journal of Educational Measurement. 2013;50(1):1-73.
32. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166. e7-. e16.
33. Lineberry M, Hornos E, Pleguezuelos E, Mella J, Brailovsky C, Bordage G. Experts' responses in script concordance tests: a response process validity investigation. Med Educ. 2019;53(7):710-722.
34. Wan SH, Tor E, Hudson JN. Commentary: expert responses in script concordance tests: a response process validity investigation. Med Educ. 2019;53(7):644-6.
35. Pau A, Thangarajoo S, Samuel VP, Wong LC, Wong PF, Matizha P, et al. Development and validation of a script concordance test (SCT) to evaluate ethical reasoning ability among first and fifth year students in a medical school. J Acad Ethics. 2019;17(2):193-204.

Appendix 1.

Examples of students' free-text explanations (direct quotes) of their clinical reasoning concerning the full or partial credit responses in each Category (A to F)

1.1. Example of Category A/B responses: a full or partial credit response from students with clinical reasoning in concordance with the panel.

Clinical scenario

A 25-year-old man presents to the Emergency Department with chest pain and shortness of breath. On examination, his BP is 110/90 mmHg, pulse 120/min.

If you were thinking of the diagnosis:	And then you find:	The diagnosis is:	
Pulmonary embolism	His left leg is swollen with dilated veins.	A B C D E -2 -1 0 +1 +2	-2 = much less likely -1 = less likely 0 = neither less nor more likely +1 = more likely +2 = much more likely

The majority of the experts in the panel had chosen a 'much more likely' response as the presence of a unilateral swollen leg with dilated veins was highly suggestive of a deep venous thrombosis (DVT) and therefore could proceed to pulmonary embolism (PE). One student's free-text response included swollen left leg with dilated veins is suggestive of DVT. A DVT predisposes to PE, which was concordant with the experts.

✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠

1.2. Example of Category C response: clinical reasoning not in concordance with the expert panel despite a full credit response from the student.

Clinical scenario

A 45-year-old presents to the Emergency Department with a 3-day history of epigastric pain.

If you were thinking of ordering:	And then you find:	The investigation is:	
ECG	The pain radiates to his left shoulder	A B C D E -2 -1 0 +1 +2	-2 = much less appropriate -1 = less appropriate 0 = neither less nor more appropriate +1 = more appropriate +2 = much more appropriate

`Expert panel's most frequently selected (modal) response was 'much more appropriate' as the patient would likely be having cardiac angina type of pain and therefore an ECG would be indicated to rule in or rule out cardiac ischaemia.

However, a Year 4 student thought that 'the pain radiating to the left shoulder is suggestive of diaphragmatic involvement, which could be due to pericarditis'; and therefore chose 'much more appropriate'; which was an incorrect clinical concept as pericarditis does not typically result in shoulder pain nor involve the diaphragm anatomically.

✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠

1.3. Example of Category D response: Student response that received no credit with clinical reasoning not concordant with an expert panel

Clinical scenario

A 64-year-old man presents with an episode of jaundice. He has denied any discomfort but is feeling itchy and lethargic.

If you were thinking of the following action:	And then you find:	The investigation is:	-2 = much less appropriate -1 = less appropriate 0 = neither less nor more appropriate +1 = more appropriate +2 = much more appropriate
Ordering a CT Abdomen	EBV IgM is elevated	A B C D E -2 -1 0 +1 +2	

The expert panel's most frequently selected (modal) response was 'much less appropriate' as the patient's jaundice was most likely due to acute EBV infection resulting in raised liver enzymes. CT abdomen was not useful in this presentation and would expose the patient to unnecessary radiations. However, a Year 4 student explained that 'EBV was often associated with gastric carcinoma and therefore a CT abdomen would be very appropriate to confirm the carcinoma in the stomach'; for which the clinical concept was incorrect.

✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠

1.4. Example of Category E response: clinical reasoning from student in concordance with the majority of the panel but the wrong response option key selected.

Clinical scenario

A 32-year-old woman presents with a 2-day history of mild cramping lower abdominal pain and light vaginal bleeding. Her last normal menstrual period was 6 weeks ago.

If you were thinking of the diagnosis:	And then you find:	The diagnosis is:	-2 = much less likely -1 = less likely 0 = neither less nor more likely +1 = more likely +2 = much more likely
Ectopic pregnancy	her serum beta HCG is 3000 IU and there is no intrauterine pregnancy identified on trans-vaginal scan	A B C D E -2 -1 0 +1 +2	

The expert panel's consensus reasoning behind the response of 'much more likely' was that a raised serum beta HCG indicated pregnancy and the transvaginal ultrasound scan (TVS) findings of the absence of intrauterine pregnancy, made the diagnosis of an ectopic pregnancy much more likely. However, a Year 3 student chose the response option of "-2" (much less likely) with a free text entry of 'beta HCG positive makes pregnancy likely, and none identified on TVS makes ectopic much more likely'; which was the consensus clinical reasoning. The student had most likely clicked the wrong key response inadvertently.

✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠ ✠

1.5. Example of a Category F response: Clinical reasoning from student in concordance with the majority of the panel but response received no credit because none of the experts had selected that particular answer option.

Clinical scenario

A 45-year-old man presents to the Emergency Department with a 3-day history of epigastric pain.

If you were thinking of the following action:	And then you find:	The investigation is:	-2 = much less appropriate -1 = less appropriate 0 = neither less nor more appropriate +1 = more appropriate +2 = much more appropriate
Ordering a Chest X-ray	bronchial breathing and crackles on right lower chest	A B C D E -2 -1 0 +1 +2	

The expert panel's unanimous response was 'much more appropriate (+2)' as the clinical signs were typical of lobar pneumonia and therefore, a chest X-ray would be much more appropriate in this clinical setting. As a result, no expert in the panel chose 'more appropriate (+1)', and this answer key, as well as the rest (0, -1, -2), did not attract any mark in the item using the classical aggregated scoring method. However, a few student participants had the appropriate clinical reasoning explanation in the free text, i.e. diagnosing lobar pneumonia, and chose 'more appropriate (+1)' to order the chest X-ray as the investigation, hence scoring a zero score for the question.

Synopsis of Chapter 7

A review of students' written think-aloud free text entries showed that the majority of students' answers to SCT items, which attracted full or partial credit, were derived based on correct clinical reasoning. These findings of very high true positive and true negative rates in students' SCT scores are evidence to support the response process validity of SCT scores. The think-aloud approach in answering SCT could be a very useful tool to enhance students' learning and understanding of their underlying thought process especially in the formative practice environment. Further collaborations with other institutions researching into the underlying thought processes of clinical reasoning using this approach is also recommended.

Chapter 8: Discussion of findings, limitations, future directions and conclusions

8.1 Discussion

Clinical reasoning skill is an essential competency for medical students, junior fellowship trainees and clinicians. It allows focused collection of relevant clinical information from the patient, accurate diagnosis, appropriate investigations and evidence-based management. Effective clinical reasoning has been shown to reduce critical medical errors and improve patient outcomes. (3, 6, 7) Accurate assessment of clinical reasoning at the undergraduate and post-graduate levels is essential to ensure competent performance. In the past decades, there have been various assessment modalities attempting to assess this pivotal skill. One of the main challenges has been the multiple plausible threats to the validity of SCT scores.

This dissertation is a systematic accumulation of a series of practitioner inquiry studies into the use of SCT in undergraduate medicine, particularly focusing on the validity of SCT scores. It aimed to address the recent call by leaders in the field of SCT for further research on the two least studied sources of evidence of validity, namely response processes and consequences. (62) The first study has demonstrated that lower performing students tended to avoid extreme responses (‘-2’ or ‘+2’ options) in answering SCT questions. Thoughtful design and balance of SCT items, however, may help to mitigate some of the validity threats to medical students’ SCT scores. The second study in this PhD project has responded to some of the published concerns about the validity threats to SCT scores where students could potentially game the examination by avoiding the extreme responses (‘-2’ or ‘+2’) or purposefully choosing the median responses (‘0’ option). (52) It also considered the issue of concern over the faulty logic of aggregated scoring in SCT. The validity by design approach seemed to be effective in mitigating the plausible threats to the validity of SCT scores, due to the examinee test-taking strategy. Validity by design was achieved through careful

selection of items in each SCT paper to ensure there was a balanced distribution of expert reference panel modal answers, i.e. in both extreme and median response options. This was an additional part of the test optimisation process for the development of each SCT paper. This paper added a structured protocol to the current literature, to improve SCT validity and address some of the potential threats to the validity of the SCT scores. However, when using such an optimising process during examination construction, one should maintain the rigour of the content validity and the characteristic of uncertainty in the clinical scenario for substantive validity. Collaboration with other medical schools to share the pool of SCT items and increase coverage of core disciplines would ensure high content validity of the examination and alignment with the construct of interest.

The third study in this PhD project, comparing the scores of novices versus experienced clinicians, showed progression in SCT scores from students in the penultimate to the final year of the medical programme, and from undergraduate students to practising GP clinicians. This study provides further response process validity evidence for SCT. (62) Some earlier studies had looked at score progression in a number of learners in different specialties. In 2009, a Canadian study used a 90-item radiation oncology SCT to assess medical students, residents and radiation oncologists. They showed statistically significant progression of scores from students to specialist. (63) Another study found that Otorhinolaryngology residents had a significantly higher web-based SCT scores than medical students. (64) A Brazilian study showed that SCT scores were higher in senior compared to junior medical students. (65) A few research studies have looked at progression in scores with postgraduate trainees in Paediatrics, Psychiatry and General Practice, (45, 66-68) and a recent study with a relatively small number of medical students showed significant progression of SCT scores from students to residents in Urology. (69) More research looking into the longitudinal progression of clinical reasoning abilities in the undergraduate setting and post-graduate fellowship training environment, stratifying according to the years of study/training, would provide further evidence of the construct validity of SCT.

The acceptable reliability of SCT scores (Cronbach alpha: 0.62 – 0.86) in this third study also provided internal structure evidence as part of SCT scores validity. In relation to content specificity, which may pose a potential threat to the validity of decisions that are based on SCT scores aggregated from items covering multiple disciplines. These decisions may be made on the assumption that SCT scores are measuring a unidimensional construct, i.e. the 'global' clinical reasoning ability that is independent of the specific contents (disciplines) in the SCT items in a test. However, in the study context, the SCT scores are not used independently to determine whether a student would pass or fail in a year-long course nor treated as a separate item within the rules for progression. Instead, SCT scores are aggregated with scores from other components of the written papers. Thus content specificity is not relevant to consideration of the validity of SCT score (use in the medical programme), and was not addressed in the research aims.

Further support for the response process validity of SCT came from the finding that the majority of the students' clinical reasoning was correctly aligned with the expert panel's consensus in clinical reasoning. Rather than just looking at medical students' keyed responses on the 5-point Likert scale, use of the 'think-aloud' approach to understand the reasoning behind student responses to SCT items, added another layer of understanding to the underlying thought process of the students. This study (Chapter 7) asked the students to justify their responses during the online tests while they were fully engaged in the thought process. In fact, a few students actually reported in their free text answers that they had employed the test-taking strategy to try to avoid the extreme responses and answered on the 'safe' side and chosen '-1' or '+1'. This unpublished finding aligns with the findings reported in the Chapter 3 paper.

The study presented in Chapter 7 also provided further evidence for the consequential validity of SCT. The short post-test evaluation survey of student perspectives of the think-aloud approach revealed that students found this approach in the formative SCT useful for supporting their learning in clinical reasoning. Other reports in the literature have supported the education impact of the 'think-aloud' approach on clinical reasoning. The written 'think-aloud' approach to explore students' response process

had previously built confidence in SCT as a valid format to test the clinical reasoning of nursing students and paediatrics trainees. (60, 70) Adding the 'think-aloud' approach to the student answering process provided medical students with enhanced post-assessment feedback opportunities, an approach that could be incorporated into regular formative assessment sessions for medical undergraduates or fellowship trainees. A recent publication from a Victorian medical school, exploring medical students' perceptions of the educational impact of SCT for learning clinical reasoning, has reported a similar conclusion. (71) Provision of constructive feedback to students or trainees by revealing the expert panel's clinical reasoning, followed by facilitated discussions, is likely to enhance learning.

In summary, the published papers presented in the thesis responded to international calls for further research into the construct validity of SCT scores. The papers have addressed the five sources of evidence for the construct validity of SCT: content; response process; internal structure; relationship to other variables; and consequences.

However, when assessing clinical reasoning we should not be focussing on only one modality of assessment. Traditional assessment programmes have tended to be based on a single instrument for the assessment of each of the domains of 'knowledge', 'skills', 'problem solving' and 'attitudes' at single time points, with potential to result in highly unwanted side effects. (23) The programmatic approach, where a student's performance is being gathered at multiple assessment points across the trajectory of the course with accompanying rich feedback to support learning, is a recommended direction for medical education. (59) With such an approach, the assessment of clinical reasoning would be derived from multiple sources and by various tools (e.g. SCT, KF and OSCE) across time. Such programmes could be ideally constructed using complementary assessment methods to account for each method's validity and feasibility issues, as well as their advantages, and disadvantages. (24)

8.2 Limitations

The findings from the research studies have limitations in that the data analysed was from just one medical school, and the set of SCT questions used for each cohort was relatively small. The research projects in the thesis have not covered the whole spectrum of sources of validity evidence for SCT scores. (72) However, as much of the SCT research has been reported in the context of postgraduate medical education, the thesis has addressed some key issues associated with the validity of SCT scores in the setting of medical undergraduate education.

Current and ongoing research collaboration with national and international health professional educators should result in findings with greater power and in additional contexts, and provide further evidence to support the use of SCT in the assessment of clinical reasoning. It could also generate ways to test the impact of enhanced feedback to students, such as the 'think-aloud' approach reported in the thesis. Use of student focus groups to understand their underlying cognitive thought processes in deciding to choose between '-2' and '-1'; or between '+2' and '+1' in the 5-point Likert scale when answering a SCT question, could shed more light on the response process validity of SCT. (73)

8.3 Ongoing research initiatives and Future directions

Validity remains a hypothesis to be tested with every set of test scores, hence validation study of scores should be an ongoing effort. As suggested in the conclusion of the first published study (Chapter 2), further investigation of the SCT 'pass-fail' cut score has been initiated. Some preliminary findings have been presented at the International Medical Education Conference in 2013 (as listed under Candidate publications & conference/workshop presentation section). These findings suggested that a cut score of 3.5 standard deviations below the expert panel's mean SCT score could be used in the undergraduate medical programme setting. Further in-depth study

of the standard-setting aspect of the validity of SCT score use and interpretation will continue.

Much of the previous research relating to SCT has been performed in the US and Canadian contexts, while the thesis has explored SCT in one Australian context. Further work should expand the use of, and research into, SCT as an assessment modality in health professional education in the Asia Pacific region. There is also the potential to respond to growing interest in the use of the SCT format in assessing medical ethics and professionalism in the field of health education. (74, 75) The inherent property of uncertainty in the question aligns with assessing medical ethics where there are usually no absolutely correct or incorrect answers. Further research into the use of SCT in this discipline could expand its application into medical ethics.

Following on from the research reported in this thesis, a variety of national and international research collaborations are currently underway. They aim to gather further evidence for the validity of SCT in assessing clinical reasoning in undergraduate and postgraduate health professional education, in the Asia Pacific region.

Ongoing and future research projects include:

- Collaboration with colleagues in the Endocrinology Department at the National University Health System, Singapore to develop new SCT questions for trainee fellows in Endocrinology. Pilot questions have been selected, and expert panel responses collated to deliver the first SCT formative examination to fellowship trainees. Research is ongoing investigating the consistency of the experts' responses and the progression of SCT scores in the various level of training in the trainees.

- Collaboration with colleagues from the School of Nursing at the National University of Singapore to develop practice SCT questions for graduate nurses

to enhance their clinical reasoning skills. The first set of SCT formative examinations will be delivered in mid-2020. Data will be collected to investigate the students' perception of the usefulness of such a tool in enhancing their clinical reasoning skills in the Nursing field.

- In collaboration with the University of Adelaide, South Australia, an online SCT and MCQ practice examination for their Year 4 & Year 5 medical students has been delivered. It incorporated the 'think-aloud' approach to enhance feedback for improving student clinical reasoning skills. Data is being analysed to look at the correlation of scores between the SCT and MCQ format for assessing clinical reasoning, exploring evidence of convergent validity of SCT scores in the relationship with other variables.

- A survey has also been administered to look into the usefulness of such a formative examination with immediate feedback on student learning, seeking further evidence for the educational impact of the SCT assessment on student learning (consequential validity evidence).

- A pilot study had been conducted in collaboration with the medical ethicist at the National University of Singapore to develop case scenarios, and use of SCT questions to engage students to learn and discuss medical ethical principles in the small group tutorial environment. The initial feedback from the students was extremely positive. A full research study in the use of the SCT format to enhance medical ethics teaching is in progress.

To further enhance this collaborative work, a research network has been set up with Asia Pacific medical educators who are interested in SCT to have a continual discussion on valuable research projects.

8.4 Conclusions

This thesis has contributed to the international literature providing further support for the validity of SCT scores in assessing clinical reasoning in undergraduate medical education. The research has shown the following: thoughtful design and balance of SCT items can mitigate some of the validity threats to medical student SCT scores; the tendency of SCT scores to progress with increasing levels of clinical practice experience is further support for the construct validity of SCT scores; and use of the 'think-aloud' approach to explore students' response process has built confidence in SCT as a valid format to test the clinical reasoning of undergraduate students, while providing them with enhanced post-assessment feedback opportunities.

By adopting the approaches investigated in the thesis research papers, namely careful design, optimisation and balancing of the items, SCT should increasingly gain acceptance in health professional education for assessing clinical reasoning. This dissertation, by supporting the construct validity of SCT scores in one setting in undergraduate medicine, as well as recommending ways to improve the validity of the tool in assessing clinical reasoning, has made a significant contribution to the current literature. It should facilitate further uptake of SCT as a valuable modality to enhance the learning and assessment of clinical reasoning. However, validation of SCT scores on each testing occasion, and in other contexts, requires ongoing effort.

References

1. Higgs J, Jones M, Loftus S, Christensen N. Clinical reasoning in the health professions. 3rd ed. Amsterdam: Butterworth-Heinemann; 2008.
2. Krairiksh M, Anthony MK. Benefits and outcomes of staff nurses' participation in decision making. J Nurs Adm. 2001;31(1):16–23.
3. Croskerry P. A universal model of diagnostic reasoning. Acad Med. 2009;84(8):1022-8.
4. Charlin B, Boshuizen HPA, Custers EJ, Feltovich PJ. Scripts and clinical reasoning. Med Educ. 2007;41(12):1178-84.
5. Alfaro-LeFevre R. Critical thinking, clinical reasoning, and clinical judgment: a practical approach. 5th ed. St. Louis, MO: Elsevier Saunders; 2013.
6. Friedman CP, Gatti GG, Franz TM, Murphy GC, Wolf FM, Heckerling PS, et al. Do physicians know when their diagnoses are correct?: Implications for decision support and error reduction. J Gen Intern Med. 2005;20(4):334-9.
7. Norman G. Dual processing and diagnostic errors. Adv Health Sci Educ Theory Pract. 2009;14 Suppl 1(S1):37-49.
8. Harasym PH, Tsai T-C, Hemmati P. Current trends in developing medical students' critical thinking abilities. Kaohsiung J Med Sci. 2008;24(7):341–55.
9. Lateef F. Clinical reasoning: the core of medical education and practice. Intern Emerg Med. 2018;1(2):1015–21.
10. Stempsey WE. Clinical reasoning: new challenges. Theor Med Bioeth. 2009;30(3):173-9.
11. Kassirer JP. Diagnostic reasoning. Ann Intern Med. 1989;110(11):893-900.
12. Young ME, Dory V, Lubarsky S, Thomas A. How different theories of clinical reasoning influence teaching and assessment. Acad Med. 2018;93(9):1415.
13. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2005;39(1):98-106.
14. Førde R. Competing conceptions of diagnostic reasoning – Is there a way out? Theor Med Bioeth. 1998;19(1):59-72.
15. Schön DA. The reflective practitioner: how professionals think in action. London: Routledge Ltd; 2017.
16. Farhan Bhanji KL, Mark Goldszmidt, Mark Walton, Kenneth Harris. CanMEDS Framework: Royal College of Physicians and Surgeons of Canada. [internet] 2017 [cited 2020 March 28]. Available from: http://www.royalcollege.ca/rcsite/canmeds/framework/canmeds-role-medical-expert-e.
17. ACGME. ACGME Common Program Requirements. [internet] 2017 [cited 2020 March 28]. Available from: http://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/CPRs_2017-07-01.pdf.
18. van der Vleuten CP, Newble DI. How can we test clinical reasoning? Lancet. 1995;345(8956):1032-4.
19. Groves M, Scott I, Alexander H. Assessing clinical reasoning: a method to monitor its development in a PBL curriculum. Med Teach. 2002;24(5):507-15.
20. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ. 2005;39(12):1188-94.
21. Thampy H, Willert E, Ramani S. Assessing Clinical Reasoning: Targeting the Higher Levels of thePyramid. J Gen Intern Med. 2019;34(8):1631-6.
22. Ber R. The CIP (comprehensive integrative puzzle) assessment method. Med Teach. 2003;25(2):171-6.
23. Schuwirth L. Is assessment of clinical reasoning still the Holy Grail? Med Educ. 2009;43(4):298-300.
24. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med. 2019:94(6):902–912.
25. Amini M, Moghadami M, Kojuri J, Abbasi H, Abadi AAD, Molaee NA, et al. An innovative method to assess clinical reasoning skills: clinical reasoning tests in the second national medical science Olympiad in Iran. BMC Res Notes. 2011;4(1):418.
26. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ. 1979;13(1):41.

27. Park WB, Kang SH, Myung SJ, Lee Y-S. Does Objective Structured Clinical Examinations score reflect the clinical reasoning ability of medical students? Am J Med Sci. 2015;350(1):64–7.
28. Lessing JN, Rendón P, Durning SJ, Roesch JJ. Approaches to clinical reasoning assessment. Acad Med. 2020:1.
29. Gruppen LD. Clinical reasoning: defining it, teaching it, assessing it, studying it. West J Emerg Med. 2017;18(1):4–7.
30. McDonald MB, McGregor RS, Bhatia D, Dickinson B, Maxwell E, Dawlabani N, et al. A milestones friendly clinical reasoning assessment tool. Acad Pediatr. 2013;13(4):e1–e2.
31. Fleiszer D, Hoover ML, Posel N, Razek T, Bergman S. Development and validation of a tool to evaluate the evolution of clinical reasoning in trauma using virtual patients. J Surg Educ. 2018;75(3):779–86.
32. Kreiter CD. A Bayesian perspective on constructing a written assessment of probabilistic clinical reasoning in experienced clinicians: Assessing clinical reasoning. J Eval Clin Pract. 2017;23(1):44-8.
33. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. BMC Med Educ. 2007;7(1):49.
34. Palmer EJ, Duggan P, Devitt PG, Russell R. The modified essay question: Its exit from the exit examination? Med Teach. 2010;32(7):e300-e7.
35. Pham H, Trigg M, Wu S, O'Connell A, Harry C, Barnard J, et al. Choosing medical assessments: does the multiple-choice question make the grade? Educ Health (Abingdon). 2018;31(2):65-71.
36. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The script concordance test: a tool to assess the reflective clinician. Teach Learn Med. 2000;12(4):189–95.
37. Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, et al. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. Med Educ. 2006;40(9):848-54.
38. Mehay R, editor. The essential handbook for GP training and education. Portland: Ringgold, Inc; 2012.
39. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 Suppl):S63-7.
40. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: From theory to practice: AMEE Guide No. 75. Med Teach. 2013;35(3):184-93.
41. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing (2014 Edition) 2020 [Available from: https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition.]
42. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7.
43. Young M, St-Onge C, Xiao J, Vachon Lachiver E, Torabi N. Characterizing the literature on validity and assessment in medical education: a bibliometric study. Perspect Med Educ. 2018;7(3):182-91.
44. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166.e7–16.
45. 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 [abstract]. Acad Pediatr. 2016;16(6):e6.
46. Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, Vleuten Cvd. Stability of clinical reasoning assessment results with the Script Concordance test across two different linguistic, cultural and learning environments. Med Teach. 2002;24(5):522-7.
47. Claessens YE, Wannepain S, Gestin S, Magdelein X, Ferretti E, Guilly M, et al. How emergency physicians use biomarkers: insights from a qualitative assessment of script concordance tests. Emerg Med. 2014;31(3):238-41.
48. Dawson T, Comer L, Kossick MA, Neubrander J. Can script concordance testing be used in nursing education to accurately assess clinical reasoning skills? J Nurs Educ. 2014;53(5):281–6.
49. Deschênes M-F, Charlin B, Gagnon R, Goudreau J. Use of a script concordance test to assess development of clinical reasoning in nursing students. J Nurs Educ. 2011;50(7):381-7.
50. Dumas JP, Blais JG, Charlin B. Script concordance test: can it be used to assess clinical reasoning of physiotherapy student? Physiotherapy. 2015;101:e332-e333.

51. Morris A, Campbell DE. The script concordance test for clinical reasoning in paediatric medicine: Medical student performance and expert panel reliability. Focus on Health Professional Education. 2015;16(2):4-12.
52. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47(12):1175-83.
53. See KC, Tan KL, Lim TK. The script concordance test for clinical reasoning: re-examining its utility and potential weakness. Med Educ. 2014;48(11):1069-77.
54. Lubarsky S, Dory V, Meterissian S, Lambert C, Gagnon R. Examining the effects of gaming and guessing on script concordance test scores. Perspect Med Educ. 2018;7(3):174-81.
55. Wan SH. Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine. Hong Kong Med J. 2015;21(5):455-61.
56. 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. 2017;18(2):26-35.
57. Wan MS, Tor E, Hudson JN. Improving the validity of script concordance testing by optimising and balancing items. Med Educ. 2018;52(3):336-46.
58. Wan MSH, Tor E, Hudson JN. Construct validity of script concordance testing: progression of scores from novices to experienced clinicians. Int J Med Educ. 2019;10:174.
59. Wan SH, Tor E, Hudson JN. Commentary: expert responses in script concordance tests: a response process validity investigation. Med Educ. 2019;53(7):644-6.
60. Power A, Lemay J-F, Cooke S. Justify your answer: the role of written think aloud in script concordance testing. Teach Learn Med. 2017;29(1):59–67.
61. Lineberry M, Hornos E, Pleguezuelos E, Mella J, Brailovsky C, Bordage G. Experts' responses in script concordance tests: a response process validity investigation. Med Educ. 2019;53(7):710–22.
62. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance testing: a review of published validity evidence. Med Educ. 2011;45(4):329-38.
63. Lambert C, Gagnon R, Nguyen D, Charlin B. The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. Radiation Oncology. 2009;4(1):7.
64. Kania RE, Verillaud B, Tran H, Gagnon R, Kazitani D, Tran Ba Huy P, et al. Online script concordance test for clinical reasoning assessment in Otorhinolaryngology: the association between performance and clinical experience. Arch Otolaryngol Head Neck Surg. 2011;137(8):751-5.
65. Roberti A, Roberti MdRF, Pereira ERS, Costa NMdSC. Script concordance test in medical schools in Brazil: possibilities and limitations. Sao Paulo Med J. 2016;134(2):116-20.
66. Carrière B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a Script Concordance Test. Ann Emerg Med. 2009;53(5):647-52.
67. Kazour F, Richa S, Zoghbi M, El-Hage W, Haddad FG. Using the script concordance test to evaluate clinical reasoning skills in psychiatry. Acad Psychiatry. 2017;41(1):86–90.
68. Subra J, Chicoulaa B, Stillmunkes A, Mesthe P, Oustric S, Bugat MER. Reliability and validity of the script concordance test for postgraduate students of general practice. Eur J Gen Pract. 2017;23(1):209-14.
69. Nazim SM, Talati JJ, Pinjani S, Biyabani SR, Ather MH, Norcini JJ. Assessing clinical reasoning skills using Script Concordance Test (SCT) and extended matching questions (EMQs): A pilot for urology trainees. J Adv Med Educ Prof. 2019;7(1):7-13.
70. Tedesco-Schneck M. Use of script concordance activity with the think-aloud approach to foster clinical reasoning in nursing students. Nurse Educ. 2019;44(5):275–7.
71. Bacchi S, Tan Y, Chim I, Dabarno M, Lubarsky S, Duggan P. Script concordance test examinations: Student perception and study approaches. Focus on Health Professional Education. 2019;20(2):75-87.
72. Crooks TJ, Kane MT, Cohen AS. Threats to the Valid Use of Assessments. Assessment in Education: Principles, Policy & Practice. 1996;3(3):265–86.
73. Gawad N, Wood TJ, Cowley L, Raiche I. The cognitive process of test takers when using the script concordance test rating scale. Med Educ. 2020;54(4):337–47.
74. Pau A, Thangarajoo S, Samuel VP, Wong LC, Wong PF, Matizha P, et al. Development and validation of a script concordance test (SCT) to evaluate ethical reasoning ability among first and fifth year students in a medical school. Journal of Academic Ethics. 2019;17(2):193–204.
75. Tsai TC, Chen DF, Lei SM. The ethics script concordance test in assessing ethical reasoning. Med Educ. 2012;46(5):527-.

Awards and Grants

1. **First Prize, Best Oral presentation. Wan SH**, Canalese R, Frost G. Using script concordance testing as a new modality of assessment for graduate entry medical students – a pilot study. Asian Medical Education Association conference (AMEA). 2011, Malaysia.
2. **First Prize in Best selected Poster. Wan SH**, Clarke R. Using a clinician panel to set the borderline mark for Script Concordance Testing (SCT) to assess clinical reasoning for graduating medical students. International Medical Education Conference (IMEC). 2013.
3. **Merit Award for Best Oral Poster Presentation**. **Wan SH**. Standard setting the borderline pass mark for Script Concordance Testing (SCT) to assess clinical reasoning – 5 years' experience. Asia Pacific Medical Education Conference (APMEC). Jan 2016.

Candidate publications & conference/workshop presentations

National and International First Author Conference Presentations

1. **Wan SH**, Canalese R, Frost G. Using script concordance testing as a new modality of assessment for graduate entry medical students – a pilot study. **First Prize, Best Oral presentation**. Asian Medical Education Association conference (AMEA). Feb 2011. Malaysia.
2. **Wan SH**, Clarke R. Using a clinician panel to set the borderline mark for Script Concordance Testing (SCT) to assess clinical reasoning for graduating medical students. International Medical Education Conference (IMEC) 2013, **1st Prize in Best selected Poster**.
3. **Wan SH**. Using Script Concordance Testing (SCT) to assess clinical reasoning: the progression from novice to practising general practitioners. Asia Pacific Medical Education Conference (APMEC). Jan 2014 (oral presentation).
4. **Wan SH**. Standard setting the borderline pass mark for Script Concordance Testing (SCT) to assess clinical reasoning – 5 years' experience. Asia Pacific Medical Education Conference (APMEC). Jan 2016. **Merit Award for Best Oral Poster Presentation**.
5. **Wan SH**, P Devitt. Lowest quartile candidates adopt test-wise strategies in answering Script Concordance Test questions. Ottawa Australia & New Zealand Association for Health Professional Educators (ANZAHPE) Conference. March 2016. Oral presentation.
6. **Wan SH**. The usefulness of formative Script Concordance Test questions in improving clinical reasoning in graduate-entry clinical year students. Asia Pacific Medical Education Conference (APMEC). Jan 2017. Oral presentation.
7. **Wan SH**, Tor E, Hudson JN. Improving the validity of Script Concordance Testing (SCT) by better item selection pre-examination. Australian & New

Zealand Association for Health Professional Educators (ANZAHPE) Conference. July 2017. Oral Presentation.

8. **Wan SH**, Tor E. Correlation of Script Concordance Test (SCT) with other assessment modalities in a graduate-entry medical course. Ottawa Medical Education Conference, March 2018. Poster presentation.
9. **Wan SH**, Tor E, Hudson JN. Construct validity of Script Concordance Test (SCT) in assessing clinical reasoning – progression from novice to general practitioner. Australia & New Zealand Association for Health Professional Educators (ANZAHPE) Conference. July 2019. Oral presentation.
10. **Wan SH**, Tor E, Hudson JN. Using the new "Think-aloud" method in Script Concordance Test to better assess Clinical Reasoning in medical students. AMEE. August 2019. Oral presentation.
11. **Wan SH**, Tor E, Hudson JN. Construct validity of Script Concordance Testing scores: progression of scores in senior medical students – a six-year follow-up. APMEC. Jan 2020, Singapore. Oral presentation.
12. **Wan SH**, Tor E, Hudson JN. Students' perception of the think-aloud approach in Script Concordance Test (SCT) to assess clinical reasoning. Ottawa Conference. March 2020, Malaysia. Oral presentation.

Invited Workshop and Presentations

1. *Using Script Concordance Test to assess clinical reasoning in the health care professions: hints and pitfalls*. Asia Pacific Medical Education Conference (APMEC). Jan 2017. Invited for running the Pre-conference Workshop with 35 health professional participants from Singapore and internationally attended.
2. *Joint workshop: Script Concordance Testing – an efficient way to assess clinical reasoning*. Hong Kong. Aug 2017. Invited by the Chinese University of Hong Kong Office of Medical Education to deliver the workshop to academics in Hong Kong.
3. *Egypt Assessment Leaders' Workshop*. Cairo. Oct 2017. Invited by the Egyptian Government Medical Education Department to run the 2-day Assessment workshop highlighting the use of Script Concordance to assess clinical reasoning.

4. *Using Script Concordance Test to assess clinical reasoning, ethics and professionalism in the health care profession*. Asia Pacific Medical Education Conference (APMEC). Jan 2019. Invited to run the Pre-conference Workshop for 25 health professional participants from Singapore and other international delegates.
5. *Using Script Concordance Test (SCT) to improve clinical reasoning in medical students.* Lee Kong Chian School of Medicine. Jan 2019, Singapore. Invited to present the use of SCT to assess students formatively and summatively.
6. *NUS Nursing: Using SCT for formative feedback in nursing presentation*. Alice Lee Centre for Nursing Studies. National University of Singapore. Singapore. April 2019.
7. *Using formative Script Concordance Test (SCT) with feedback to improve clinical reasoning in medical students.* August 2019, Perth. Invited workshop at University of Western Australia.
8. *Enhancing clinical reasoning by formative Script Concordance Test.* Brisbane April 2020. Invited delivery of the workshop at Griffith University.

Appendices

Appendix 1: Statement of Contribution by Others

Title: ***Literature Review: Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine***

Author

S. H. Wan

Journal

Hong Kong Medical Journal (**ERA Journal**). 2015(21):455-61.

Author contribution

SHW (candidate) is the principal and sole author of the paper. He conducted the literature review, rewriting multiple drafts of the manuscript, acted as the corresponding author responding to reviewer's reports and coordinating submission and publication of the manuscript.

Author signatures:

Name	**Date**
S. H. Wan	15/04/2020

Appendix 2: Statement of Contribution by Others

Title: ***Association between candidate total scores and response pattern in script concordance testing of medical students***

Authors

S. H. Wan, P. Duggan, E. Tor & J. N. Hudson

Journal

Focus on Health Professional Education: A Multi-disciplinary Journal (**ERA Journal**). 2017;18(2):26-35.

Author contributions

SHW (candidate) is the principal author of the paper. He conceived the idea and designed the study, collect and analysed all data, rewriting all drafts of the manuscript, acted as the corresponding author responding to reviewer's reports and coordinating submission and publication of the manuscript.

PD reviewed and comment on the methodology of the manuscript.

ET supervised the research and assisted in the statistical analysis of the data, reviewed and commented on all drafts of the manuscript.

JNH supervised the research, assisted with reviewing the current literature, reviewed and commented on all drafts of the manuscript.

Author signatures:

Name	**Date**
S. H. Wan	15/04/2020
P. Duggan	
E. Tor	
J. N. Hudson	

Appendix 3: Statement of Contribution by Others

Title: ***Improving the validity of script concordance testing by optimising and balancing items***

Authors

S. H. Wan, E. Tor & J. N. Hudson

Journal

Medical Education (ERA Journal). 2018;52(3):336-46

Author contributions

SHW (candidate) is the principal author of the paper. He conceived the idea and designed the study, optimise the selection and balancing of exam items, collect and analysed all data, revising all drafts of the manuscript, acted as the corresponding author responding to reviewer's reports and coordinating submission and publication of the manuscript.

ET supervised the research and assisted in the statistical analysis of the data, contributed to the thinking and critique of the evaluation and manuscript; and commented on all drafts of the manuscript.

JNH supervised the research, assisted with reviewing the current literature and reviewed and commented on the manuscript.

Author signatures:

Name	**Date**
S. H. Wan _____	15/04/2020
E. Tor _____	
J. N. Hudson _____	

Appendix 4: Statement of Contribution by Others

Title: ***Construct validity of Script Concordance Testing scores: progression from medical students to general practitioners***

Authors

S. H. Wan, E. Tor & J. N. Hudson

Journal

IJME (ERA Journal). 2019;10:174.

Author contributions

SHW (candidate) is the principal author of the paper. He conceived the idea and designed the study, developed the test items, collect and analysed all cohort data, revising all drafts of the manuscript, acted as the corresponding author responding to reviewer's reports and coordinating submission and publication of the manuscript.

ET supervised the research and assisted in the statistical analysis of the data, contributed to the thinking and critique of the evaluation and manuscript; and commented on the manuscript.

JNH supervised the research, assisted with reviewing the current literature and reviewed and commented on multiple drafts of the manuscript.

Author signatures:

Name	**Date**
S. H. Wan _____	15/04/2020
E. Tor _____	
J. N. Hudson _____	

Appendix 5: Statement of Contribution by Others

Title: ***Commentary: Expert responses in script concordance tests: A response process validity investigation***

Authors

S. H. Wan, E. Tor & J. N. Hudson

Journal

Medical education (**ERA Journal**). 2019;53(7):644-6.

Author contributions

SHW (candidate) is the principal author of the commentary. He revised multiple drafts of the commentary, acted as the corresponding author responding to the editor's comments.

ET and JNH contributed by giving expert comments to the original article and reviewing the commentary before final submission.

Author signatures:

Name		**Date**
S. H. Wan	_____	15/04/2020
E. Tor	_____	
J. N. Hudson	_____	

Appendix 6: Statement of Contribution by Others

Title: ***Examining response process validity of Script Concordance Testing: a think-aloud approach***

Authors

S. H. Wan, E. Tor & J. N. Hudson

Journal

IJME (ERA Journal). 2020;11:127-135.

Author contribution

SHW (candidate) is the principal author of the paper. He conceived the think-aloud approach and designed the study, conducting the online exam, collect and analysed all data, revising multiple drafts of the manuscript, acted as the corresponding author responding to reviewer's reports and coordinating submission and publication of the manuscript.

ET supervised the research and assisted in the statistical analysis of the data, contributed to the thinking and critique of the evaluation and manuscript; and commented on multiple drafts of the manuscript.

JNH supervised the research, assisted with reviewing the current literature and reviewed and commented on multiple drafts of the manuscript.

Author signatures:

Name	**Date**
S. H. Wan _____	15/04/2020
E. Tor _____	
J. N. Hudson _____	

Appendix 7: Script Concordance Testing online quiz for Year 3 (Chapter 7)

<table>
<tr><td colspan="4"><u>Clinical Scenario A</u>

A 22-year-old female medical student comes to see you complaining of fatigue, weight loss and night sweats of about one month's duration. She recently returned from her elective in India. She is a non-smoker and apart from a cardiac murmur detected in childhood there is no significant past medical history. She has no diarrhoea and no abdominal pain.</td></tr>
<tr><td>If you are considering the following investigation…</td><td>and then you find that…</td><td>you would then consider the investigation to be…</td><td rowspan="3">-2 : much less useful
-1 : less useful
0 : neither more nor less useful
+1 : more useful
+2 : much more useful</td></tr>
<tr><td>1. A CT scan of chest, abdomen and pelvis</td><td>there are splinter haemorrhages on examination</td><td>A B C D E
-2 - 1 0 +1 +2
0 1 0.4 0 0</td></tr>
<tr><td>2. An echocardiogram</td><td>there are ring-like parasites with double nuclear dots in erythrocytes on the manual blood film</td><td>A B C D E
-2 - 1 0 +1 +2
1 0.7 0 0 0</td></tr>
</table>

Clinical Scenario B

A 32-year-old woman presents with a 2-day history of mild cramping lower abdominal pain and light vaginal bleeding. Her last normal menstrual period was 6 weeks ago.

If you were thinking of...	and then you find that…	this hypothesis becomes…	
3. Ectopic pregnancy	her serum beta HCG is 3000 IU and there is no intrauterine pregnancy identified on transvaginal scan	**A B C D E** -2 - 1 0 +1 +2 0 0 0 0.7 1	**-2 :** much less likely **-1 :** less likely **0 :** neither more nor less likely **+1 :** more likely **+2 :** much more likely
4. Menstruation	she has polycystic ovarian syndrome	**A B C D E** -2 - 1 0 +1 +2 0 0 0.1 1 0	

Clinical Scenario C

A mother brings in a two-year-old child to you with fever. On examination the child is irritable, has a mild fever and the left eardrum is congested and bulging. You diagnose unilateral acute otitis media. The child has not received antibiotic treatment before.

If you are considering the following treatment or action…	and then you find that…	you would then consider the treatment or action to be...	
5. Prescribing amoxycillin	The mother had an anaphylactic reaction to penicillin	**A B C D E** -2 - 1 0 +1 +2 0 0.2 1 0 0	**-2 :** much less appropriate **-1 :** less appropriate **0 :** neither more nor less appropriate **+1 :** more appropriate **+2 :** much more appropriate
6. Arranging an assessment by an audiologist	The child has a speech delay	**A B C D E** -2 - 1 0 +1 +2 0 0 0 0.5 1	

Clinical Scenario D

A ten-year-old boy presents with acute jaundice and right upper quadrant pain. He has had a preceding sore throat.

If you are considering the following investigation…	and then you find that…	you would then consider the investigation to be…	
7. Hepatitis C antibodies	His mother had a blood transfusion for postpartum haemorrhage	**A B C D E** -2 -1 0 +1 +2 0 0 1 0 0	**-2 :** much less useful **-1 :** less useful **0 :** neither more nor less useful **+1 :** more useful **+2 :** much more useful
8. Epstein Barr virus IgM	His GP has treated his sore throat with Amoxycillin and he has a skin rash	**A B C D E** -2 -1 0 +1 +2 0 0 0 0 1	
9. Hepatitis A IgM	Some of his classmates have been unwell and have had loose bowel motions	**A B C D E** -2 -1 0 +1 +2 0 0 0.33 0.33 1	

Clinical Scenario E

A 35-year-old woman asks for your help to lose weight. She tells you that she eats large volumes of junk food late at night, or when she fights with her partner.

If you were thinking of...	and then you find that…	this hypothesis becomes…	
10. Anorexia nervosa	She has a BMI of 22 (normal 20 - 24.9)	**A B C D E** -2 -1 0 +1 +2 1 0.72 0 0 0	**-2 :** much less likely **-1 :** less likely **0 :** neither more nor less likely **+1 :** more likely **+2 :** much more likely
11. Hypothyroidism	Her resting pulse rate is 88 /min	**A B C D E** -2 -1 0 +1 +2 0.4 1 0.2 0 0	
12. Bulimia	There are macerations on two fingers of her left hand	**A B C D E** -2 -1 0 +1 +2 0 0 0 1 0.29	

Appendix 8: Script Concordance Testing online quiz for Year 4 (Chapter 7)

<table>
<tr><td colspan="4"><u>Clinical Scenario A</u>

A 65-year-old homeless man presents to the Emergency Department with fever, cough and purulent sputum for 2 days. He now complains of sharp right-sided chest pain, worse with respiration.</td></tr>
<tr><td>If you were thinking of...</td><td>and then you find that…</td><td>this hypothesis becomes…</td><td rowspan="3">-2 : much less likely
-1 : less likely
0 : neither more nor less likely
+1 : more likely
+2 : much more likely</td></tr>
<tr><td>1. Pneumothorax</td><td>There is marked dullness to percussion over the right lower lobe</td><td>A B C D E
-2 -1 0 +1 +2
1 0.1 0 0 0</td></tr>
<tr><td>2. Aspiration pneumonia</td><td>Sputum cultures are negative after two days</td><td>A B C D E
-2 -1 0 +1 +2
0 0.6 1 0.3 0</td></tr>
</table>

Clinical Scenario B

A mother presents to the Emergency Department with her 3-year-old daughter who has had a persistent cough and wheeze since last night.

If you were thinking of...	and then you find that…	this hypothesis becomes…	
3. Cystic fibrosis	The child has not had any breathing problems before this episode	**A B C D E** -2 -1 0 +1 +2 0.3 1 0.5 0 0	**-2 :** much less likely **-1 :** less likely **0 :** neither more nor less likely **+1 :** more likely **+2 :** much more likely
4. Inhalation of a foreign body	The symptoms improve after 10 minutes on a nebuliser mask	**A B C D E** -2 -1 0 +1 +2 1 0.8 0 0 0	

Clinical Scenario C

A 47-year-old woman presents with an episode of upper abdominal pain. She is not experiencing any radiation of the pain but is feeling nauseated.

If you were thinking of...	and then you find that…	this hypothesis becomes…	
5. Acute pancreatitis	She has a normal serum amylase and lipase level	**A B C D E** -2 -1 0 +1 +2 1 0.3 0.1 0 0	**-2 :** much less likely **-1 :** slightly less likely **0 :** neither more nor less likely **+1 :** slightly more likely **+2 :** much more likely
6. Peptic ulcer disease	She has a history of frequent recurrent gout	**A B C D E** -2 -1 0 +1 +2 0 0 0.5 1 0.3	

Clinical Scenario D

A 25-year-old woman presents complaining of vomiting intermittently for one week.

If you are considering the following treatment or action…	and then you find that…	you would then consider the treatment or action to be...	
7. Prescribing metoclopromide (Maxolon)	She has needed benztropine (Cogentin) in the past for a reaction to an antiemetic	**A B C D E** -2 -1 0 +1 +2 1 0.1 0 0 0	**-2 :** much less appropriate **-1 :** less appropriate **0 :** neither more nor less appropriate **+1 :** more appropriate **+2 :** much more appropriate
8. Prescribing intravenous Tropisetron hydrochloride (Navoban)	She has been taking an antidepressant	**A B C D E** -2 -1 0 +1 +2 0.1 0.6 1 0 0	

Clinical Scenario E

A 64-year-old man has been married for 20 years. He presents with an episode of jaundice. He has denied any discomfort but is feeling itchy and lethargic. You decide to order some investigations for Mr White.

If you are considering the following investigation…	and then you find that…	you would then consider the investigation to be…	
9. Abdominal ultrasound	The patient has had a melanoma removed from his scalp	**A B C D E** -2 -1 0 +1 +2 0 0 0.6 0.7 1	**-2 :** much less useful **-1 :** less useful **0 :** neither more nor less useful **+1 :** more useful **+2 :** much more useful
10. Hepatitis A serology	He has not travelled overseas in the last 12 months	**A B C D E** -2 -1 0 +1 +2 0.4 1 1 0 0	
11. CT Abdomen	EBV IgM is elevated	**A B C D E** -2 -1 0 +1 +2 1 0.9 0.3 0 0	

Appendix 9: Copyright permissions

License Number	4776260760788
License date	Feb 25, 2020
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Medical Education
Licensed Content Title	Improving the validity of script concordance testing by optimising and balancing items
Licensed Content Author	Judith Nicky Hudson, Elina Tor, Michael SH Wan
Licensed Content Date	Jan 9, 2018
Licensed Content Volume	52
Licensed Content Issue	3
Licensed Content Pages	11
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Electronic
Portion	Full article
Will you be translating?	No
Order reference number	Wan PhD 2020
Title of your thesis / dissertation	Exploring the validity of Script Concordance Testing to assess the Clinical Reasoning of medical students
Expected completion date	May 2020
Expected size (number of pages)	150
Requestor Location	Uni of Notre Dame 160 Oxford Street Darlinghurst Sydney, NSW 2010 Australia Attn: Uni of Notre Dame
Publisher Tax ID	EU826007151

License Number	4814510201039
License date	Apr 22, 2020
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Medical Education
Licensed Content Title	Commentary: expert responses in script concordance tests: a response process validity investigation
Licensed Content Author	Judith N Hudson, Elina Tor, Siu Hong Wan
Licensed Content Date	Apr 15, 2019
Licensed Content Volume	53
Licensed Content Issue	7
Licensed Content Pages	3
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Electronic
Portion	Full article
Will you be translating?	No
Title	Exploring the validity of Script Concordance Testing to assess the Clinical Reasoning of medical students
Institution name	n/a
Expected presentation date	May 2020
Order reference number	Michael SCT 2020 April
Requestor Location	Uni of Notre Dame 160 Oxford Street Darlinghurst Sydney, NSW 2010 Australia Attn: Uni of Notre Dame
Publisher Tax ID	EU826007151