2015-09-30

Development of a Method to Measure Clinical Reasoning in Pediatric Residents: The Pediatric Script Concordance Test

Cooke, Suzette

UNIVERSITY OF CALGARY

Development of a Method to Measure Clinical Reasoning in Pediatric Residents: The Pediatric Script Concordance Test

by

Suzette Rose Cooke

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MEDICAL SCIENCE

CALGARY, ALBERTA

SEPTEMBER, 2015

Abstract

Context: National medical examinations boards have requested that methods be developed to assess the clinical reasoning competency of medical trainees. The Script Concordance Test (SCT) is an emerging method of assessment that holds promise for the evaluation of clinical reasoning skills.

Objectives: This study had 3 objectives: 1) to examine the validity of SCT scores in accurately discriminating clinical reasoning ability between junior (JR) and senior (SR) pediatric residents and experienced pediatricians serving on the panel of experts (POE), 2) to determine if higher reliability of the SCT method could be achieved by applying specific strategies to the design and construction, and 3) to explore participants' acceptability of the SCT method.

Methods: A 90-minute SCT containing 24 cases and 137 questions was administered to 91 residents from four pediatric residency training centers. Twenty-one pediatricians served on the POE. All participants completed a post-test survey designed to explore their impressions and attitudes regarding the SCT assessment method.

Results: Overall, there was a difference in performance across the three levels of experience, $F = 22.84$ ($df = 2$); $p < 0.001$. The POE scored higher than both the SR (mean difference 9.15; $p < 0.001$) and the JR (mean difference = 14.90; $p < 0.001$) and the SR scored higher than the JR (mean difference = 5.76; $p < 0.002$). The reliability of the SCT scores (Cronbach's α) was 0.85. Sub-test data, based exclusively on the new evolving type cases demonstrated similar results. Participants expressed keen interest and engagement in the SCT format of assessment.

Conclusions: The findings of this study contribute to a growing body of literature suggesting that SCT holds promise as a valid, reliable and acceptable method to assess the core competency of clinical reasoning in medicine. The SCT method is based on two core principles we believe are central to the assessment of clinical reasoning: acknowledging uncertainty and the possibility of more than one "right answer." We propose the SCT may be effectively integrated into formative residency assessment and with increasing exposure, experience and refinement may soon be ready to pilot within summative assessments in pediatric medical education.

Acknowledgments

Reaching a major goal is rarely ever achieved alone. Completing this PhD was no exception.

I would like to thank the members of my PhD committee. Dr. Tanya Beran, my supervisor … your guidance, patience and unwavering support have been a constant source of direction and hope in my path. Dr. Harish Amin and Dr. Jean Francois Lemay: your mentorship throughout my career and your numerous contributions as members of this committee has been immensely valuable. Dr. Elizabeth Oddone Paolucci – many thanks for stepping up to review the final papers and assist with my defense.

I am also grateful for the support of the Royal College of Physicians and Surgeons of Canada: 2012 Recipient for a Fellowship for Studies in Medical Education. These funds made this PhD research project possible.

I would like to extend my gratitude to my medical education colleagues and friends in the Section of Hospital Pediatrics at the Alberta Children's Hospital in Calgary. You inspired me to pursue this dream and supported me throughout this entire venture. There are many of you. You know who you are.

A heartfelt thank-you to Dr. Bernard Charlin, inventor of the Script Concordance Test.
Bernard – I sincerely appreciate your willingness to share your passion and your insights.

I wish to thank my immediate and extended family members. To my husband Tim, for supporting me in following this dream. To my children, Katie and Noah: while you pursued junior and senior high school … we were all students together! To my mother Marlies and my siblings: Michael, Carla, Shelley and Renee. You have been a constant source of support and encouragement throughout my life. Helping me reach this goal has been further evidence of this.

Finally, a thank-you to my mentally disabled patient at the Children's Hospital who, upon me leaving to go to a PhD class one day asked: "Dr. Cooke, where you going?" I said, "I am going to school." She promptly replied with a highly inquisitive look, "Dr. Cooke, what grade you in? I replied (after considering all my years of grade school and university), "Grade 32." To which my patient most appropriately asked: "Dr. Cooke, what wrong with you?!"

In truth, I am striving to follow the example of Mahatma Gandhi:

"Live as if you were to die tomorrow. Learn as if you were to live forever."

This PhD thesis is dedicated to five people who have passed into the next life yet have touched my life in a very meaningful way. These people include my grandparents (Hugo and Maria Braun), my father (Paul Smith), my aunt (Helen Smith) and my uncle (Denis Braun). You were all very caring, principled and passionate people. Thank-you for sharing your gifts with me. I hope that in some small way I can emulate these qualities in my life.

Table of Contents

List of Figures

Chapter 2

Chapter 4

Chapter 6

List of Abbreviations

Script Concordance Test (SCT)

Pediatric Script Concordance Test (PSCT)

Royal College of Physicians and Surgeons of Canada (RCPSC)

American Accreditation Council for Graduate Medical Education (AAC-GME)

Multiple Choice Question (MCQ)

Short Answer Question (SAQ)

Objective Structured Clinical Examination (OSCE)

Panel of Experts (POE)

Analysis of Variance (ANOVA)

Post Graduate Year (PGY)

Chapter One

Introduction

An enormous amount of time and energy is invested in the teaching and assessment of medical trainees, all in the hopes that "at the end of the day" we are helping to support the next generation of physicians in providing the best possible care to our current and future patients. Just as in other areas of life, it can be fruitful to intermittently reflect on the question: "Is what we are producing 'matching' with what we are intending?" If the intent and the outcomes do not match, what specifically is missing? Can we identify these areas, strategize and make improvements to help draw us closer to our ultimate goal?

The purpose of my research is to address the gap identified between existing methods of assessment of clinical reasoning and clinical reasoning skills necessary to practice competently in modern day medicine. What has created this gap? At least four variables have contributed. First, and especially over the past 20 years, there has been a proliferation of scientific *knowledge* to the point where it is impossible for any individual physician to learn, remember and maintain all knowledge relevant to their field. Second, (and fortunately), the evolution of information technology has permitted relevant and up-to-date knowledge to be readily accessible. Third, patients are surviving and living longer with complex multi-system conditions. Given there are no algorithms to serve as references for each of these "patient constellations," physicians must apply and integrate information from a variety of sources to successfully manage these patients. Fourth, the complexity of medicine has evolved to a state where recognition and integration of contextual factors is essential to effective clinical decision-making. These *contextual factors* include an awareness of patient specific language, culture,

belief-systems, goals of care and “interfacing factors” such as health care system structures, policies and resources. Increasingly, legal and ethical considerations also come into play.

The cumulative effect of these factors has shifted the medical expert role from one that was predominantly focused on knowledge acquisition and retention to one that increasingly requires more selective applications, integration of multiple inputs, synthesis of a working plan and continuous evaluation of a patient’s status and response to treatment. These cognitive functions lie at the heart of clinical reasoning and decision-making. Yet, formative and summative assessment schemes in the domain of medical expertise have lagged behind and remain primarily focused on scientific knowledge, comprehension and recall. In addition, exam developers have traditionally avoided creating assessments that contain uncertainty, even though the complexity and realities of clinical medicine have always contained uncertainty. Furthermore, physicians practicing in the modern world era explain that there can be many acceptable ways to successfully manage a patient. However, in traditional forms of assessment, cases and questions are typically designed to accept only one “right answer.” The time is ripe to address “the gap” in pursuit of new forms of assessment that systematically test higher order, integrative clinical reasoning, reflect the inherent uncertainty in practice and engage the potential for more than one acceptable approach to a clinical problem.

Script concordance testing (SCT) has recently emerged as a potentially viable, feasible and engaging method designed to assess the competency of clinical reasoning in medicine. SCT is a web-based assessment that permits clinical scenarios to be presented with realistic accessories (diagnostic images, photographs, audio and video). SCT design is based on three core principles: 1) candidates are presented with clinical situations that contain some degree of

uncertainty and must choose between several realistic options, 2) the response format is intended to reflect the way information is processed in challenging clinical decision-making situations; candidates must identify and select key information and then integrate and apply it to a variety of clinical situations, 3) scoring takes into account the variability of responses of experts to these same situations. Each of these SCT design features help to address the gap that has been defined between clinical medicine and current assessment methods. SCT is gaining significant momentum. However, as with the early development of all methods of assessment, there are areas that require specific attention to aid in providing the evidence needed to support the validity, reliability, defensibility and acceptability of these assessments for the purposes of formative, and especially summative evaluations.

The purpose of the first paper is to report and discuss the overarching results of an SCT that was conducted within the context of pediatric in-patient medicine. This SCT was administered to 91 general pediatric residents from four major Canadian pediatric residency training programs and 21 staff physicians who served on the panel of experts. Based on the above gaps and needs, the first objective of this study was to examine the *validity* of SCT scores in accurately discriminating clinical reasoning ability between pediatric residents and experienced general pediatricians. The second objective was to determine if higher *reliability* of the SCT method could be achieved through specific strategies applied to the design, construction and administration of this SCT. The third objective was to seek feedback from SCT participants' as there are very limited data from these key stakeholders in previously published SCT studies. It is proposed this feedback is valuable and could contribute to improvements in both SCT design and the SCT experience itself. Furthermore, learning about the perspectives

and experiences of participants is essential if SCT is to gain acceptance in formative and summative evaluation schemes. The final goal of the overarching study was to introduce a new style of SCT case called the *evolving type case*. We proposed the evolving type case might more accurately portray the true timing and flow of decision-making in clinical practice. Chapter 3 is dedicated to a focused analysis of this subset of cases.

Chapter Two

Development of a Method to Measure Clinical Reasoning in Pediatric Residents: The Pediatric Script Concordance Test

2.1 Abstract

Context: National medical examinations boards have requested that methods be developed to assess the clinical reasoning competency of medical trainees. The Script Concordance Test (SCT) is an emerging method of assessment that holds promise for the evaluation of clinical reasoning skills. There have been some challenges with respect to the validity of SCT scores. Studies exploring the acceptability of the SCT method have also been lacking.

Objectives: This study had 3 objectives: 1) to examine the validity of SCT scores in accurately discriminating clinical reasoning ability between junior (PGY 1-2) and senior (PGY 3-4) pediatric residents and experienced general pediatricians, 2) to determine if higher reliability of the SCT method could be achieved by applying specific strategies to the design and construction and 3) to explore trainees' and practicing physicians' acceptability of the SCT method.

Methods: A 90-minute SCT test containing 24 cases and 137 questions was administered to 91 residents from four nationally accredited pediatric residency training centers. Twenty-one pediatricians served on the panel of reference (POE). Participants completed a post-test survey designed to explore participants' impressions and attitudes regarding the SCT method.

Results: Overall, there was a difference in performance across the three levels of experience, $F = 22.84$ ($df = 2$); $p < 0.001$. The POE scored higher than both the senior residents (mean difference 9.15; $p < 0.001$) and the junior residents (mean difference = 14.90; $p < 0.001$) and the senior residents scored higher than junior residents (mean difference = 5.76; $p < 0.002$). The reliability of the SCT

scores (Cronbach's α) was 0.85. Participants generally expressed keen interest and engagement in this form of assessment.

Conclusions: The findings of this study contribute to a growing body of literature suggesting that SCT holds promise as a valid, reliable and acceptable method to assess the core competency of clinical reasoning in medicine. We propose the SCT may be effectively integrated into formative residency assessment and with increasing exposure, experience and refinement may soon be ready to pilot within summative assessments in pediatric medical education.

2.2 Introduction

Competent and experienced physicians utilize clinical reasoning to process information necessary to make effective and efficient clinical decisions.[1, 2] There is an assumption that trainees gradually build clinical reasoning skills over the course of medical school and residency training.[3] There is also an expectation that when residency education is completed, physicians possess the clinical reasoning skills essential for independent medical practice. Contemporary methods of assessment, primarily testing knowledge and comprehension, include multiple choice questions (MCQ's), short answer questions (SAQ's) and objective structured clinical examinations (OSCE's). Currently, however, there is no dedicated method of assessment routinely used in either formative appraisals or certifying examinations in residency education to specifically evaluate clinical reasoning skills. Recognizing this deficiency, both the Royal College of Physicians and Surgeons of Canada (RCPSC)[4] and the American Accreditation Council for Graduate Medical Education in the United States (AAC-GME)[5] have requested a method be developed to assess the clinical reasoning competency of medical trainees. The Script Concordance Test (SCT) is an emerging method of assessment that holds

promise for the evaluation of clinical reasoning skills.[6] Lubarsky et al. (2013) have conducted a comprehensive review of the SCT method.[7]

Any newly proposed method of assessment must meet specific criteria to be considered worthy of integration into formative and especially summative examinations. The assessment must have strong evidence of validity, reliability, feasibility and acceptability. Over the past decade, researchers have been studying the psychometrics of the SCT assessment method. Growing evidence suggests that well-written SCTs can achieve excellent construct validity (extent to which SCT accurately measures clinical reasoning); however, studies have inconsistently shown discriminant validity (higher scores for those more experienced) within different levels of a training group.[8-10] There has also been significant recent debate about SCT and response score validity including concerns that the simple avoidance of extreme responses on the Likert scale could increase test scores.[11,12] SCT research has also revealed some inconsistency in reliability with scores ranging between 0.40 and 0.90.[13-18] These inconsistencies may be at least partly influenced by small sample sizes, heterogeneous trainees within the same study, sub-optimal combinations of cases and questions and inconsistent standards used for test development and scoring. Finally, only a few studies have purposefully examined the acceptability of this new assessment method from the point of view of trainees and practicing physicians.[8,9,16] If SCT is to be seriously considered for future formative and summative assessments, it is critical to gain this insight.

Based on the above gaps and needs, the **first** objective of this study was to examine the ***validity*** of SCT scores in accurately discriminating clinical reasoning ability between junior (PGY 1-2) and senior (PGY 3-4) pediatric residents and experienced general pediatricians. The **second** objective was to determine if higher ***reliability*** of the SCT method could be achieved through an adequate sample

size of residents, clearly defined SCT content, an optimal combination of cases and questions and consistent standards for scoring. It was proposed that these outcomes may help inform whether SCT can meet the reliability standards necessary for utilization as: 1) a method of assessing clinical reasoning in annual *formative* assessments over the course of a residency training program (Cronbach's α reliability coefficient of 0.7 or higher) and 2) a unique measurement of clinical reasoning (within the CanMEDS medical expert role) in specialty *qualifying* examinations (Cronbach's α reliability coefficient of 0.80 or higher). A **third** objective was to explore trainees' and practicing physicians' impressions and attitudes about the SCT method and whether or not they would support the incorporation of SCT into future strategies of trainee assessment.

2.3 Methods

PSCT Design

The Pediatric Script Concordance Test (PSCT) was constructed by three RCPSC certified general pediatricians, each of whom had a minimum of seven years of clinical in-patient experience, possessed formal training and experience in test development and were familiar with SCT format and methodology. The PSCT was designed using the guidelines for construction as described by Fournier et al.[19] A PSCT "test blueprint" was developed using the RCPSC Pediatrics "Objectives of Training."[20] Cases and questions were intentionally created to: a) ensure a wide array of clinical cases typical of general pediatric in-patient medicine, b) target the three primary clinical decision-making situations: diagnosis, investigation and treatment, c) contain varying levels of uncertainty to accurately represent real life clinical decision-making and d) reflect varying degrees of difficulty to appropriately challenge trainees across a four-year training spectrum.

Two different styles were utilized in the PSCT case presentations. These were designated as "classical" and "evolving" style cases. Classical cases followed the traditional SCT case design[6] with a stem followed by a series of "if you were thinking "x" and then you learn "y", the likelihood of the impact is "z" (See Figure 1). We created and introduced a new type of SCT case style called the "evolving case." We propose the evolving case style may better reflect the true sequential timing and flow of medical practice. The evolving case style follows a two-step approach where an initial scenario and possible diagnoses are presented. The participant is asked about the likelihood of each diagnosis. In the second step, new (subsequent) information about the case is presented (a second stem) and the participant is then asked to evaluate the impact this specific information has on the original set of diagnoses or on an investigation or treatment decision for the same case (see Figure 2).

Approval for this study was sought and obtained from research ethics boards at each of the four respective university study sites. A web-based design was utilized to administer the PSCT.[21] This web-based test format combined with a pre-loaded USB stick permitted the integration of audio (heart sounds), visual images (x-rays, rashes, a growth chart and an ECG) and video (a child with respiratory distress and an infant with abnormal movements) into the PSCT. It was proposed that this test design could more closely simulate real life situations in pediatric in-patient medicine.

Raw Scores

Resident responses to each question were compared with the aggregate responses of the panel of experts as described by Fournier et al.[19] All questions on the PSCT were equally weighted and had the same maximum (1) and minimum (0) values. The sum of scores for SCT questions provided the final raw score for each participant.

Standardization and Score Transformation

Standardization is a way to express deviation from a mean within a distribution of scores. Typical methods of standardization express deviation from the mean of all test takers. For an SCT however, the final score is intended to reflect the score between examinees and the panel of reference. Accordingly, scores for PSCT panel members were computed by treating each member independently as an examinee. That individual is then scored against the remaining panel members. Next, each of the panelist's scores was converted into a scale based on the mean and standard deviation of the panelists (not the mean and standard deviation of all test-takers). The mean of the panel therefore served as a reference value and the standard deviation of the panel was a gauge by which examinee performance was measured.

Based on the principles outlined above, score transformation for the examinees (residents) was performed in a two step process as outlined by Charlin et al.[17] In step one, *z* scores were calculated with a mean and standard deviation of the panel set at 0 and 1 respectively. In step two, *z* scores were transformed to *T* (final) scores by setting the panel mean and standard deviation at 80 and 5, respectively. These scores reflect an expected mean score out of 100%, thereby allowing participant scores to be easily compared.

Participants

RCPSC certified general pediatricians who were currently in practice and possessed a minimum of three years of full-time clinical experience in pediatric in-patient medicine were recruited from the local site to serve on the panel of experts (POE). Pediatric residents (postgraduate years 1-4) from four universities in Western Canada were recruited to participate in the study. The study was

introduced in person to staff (during a monthly meeting) and to residents (during academic half-day) by the primary investigator at the local site, by video teleconference and slide presentation to two of the sites, and, by a local staff presenter at the fourth site. Both groups received an orientation to the PSCT format and the classical and evolving cases. An email invitation followed each presentation. Recruitment occurred within two months of data collection. Each participant provided written consent prior to test administration.

PSCT Pilot and Optimization

The PSCT was piloted with three residents and two staff members to assess: a) test content and duration and b) technical feasibility. Test content included test readability, perceived interpretation of cases and questions, and, perceived difficulty. The latter items were measured by means of a post-test written survey. Pilot test duration times were recorded. Technical feasibility included: a) maintenance of the Internet connection to the web-based site and, b) perceived ease of navigation between USB accessories and the PSCT web cases. The information obtained from the pilot served as the basis for optimization of PSCT cases and questions.

The pilot version of the PSCT consisted of 31 cases and 186 questions. A total of 7 cases and 49 questions were removed for the following reasons: two cases were found to have multiple interpretations, two cases were deemed to be excessively long or complex, one case was judged too easy and two cases were removed to reduce test length. The optimized (final) version of the PSCT consisted of 24 cases and 137 questions.

PSCT Administration

The PSCT was administered to the panel of experts followed by administration to pediatric residents during their academic half-day at each of the four university sites over a five-week period in February and March, 2013. The co-investigator and a research assistant supervised all test administrations. Each testing session began with a 20-minute orientation including: 1) an orientation to the session (welcome and thank-you to participants, introduction of research personnel, session agenda), 2) a summary of the SCT concept and on-line testing format, 3) a review of the classical and evolving type cases, 4) a reminder about the test scope (acute care in-patient general pediatrics), test scale (number of cases and questions) and target test time (90 minutes), and, 5) instructions for navigation between the PSCT website and the USB stick. Each participant independently completed the PSCT. The web-based program tracked individual responses during the test in "real time." Test administrators also tracked completion times. Participants who had not yet completed the PSCT by 90 minutes were identified and the last question completed by the 90-minute mark recorded. While all participants were encouraged to complete the test (and did so), their final score was calculated based on responses received by the 90-minute mark.

The PSCT was followed by a 10-minute post-test web-based survey. This brief survey invited the panel of experts and all residents to provide feedback on the PSCT examination experience for six variables: 1) similarity of cases with real life clinical problems, 2) perceived representativeness of pediatric acute care medicine, 3) likability of SCT as a method of assessment, 4) degree to which the PSCT covered a range of difficulty, 5) preference for classical vs. evolving style cases and 6) perceived utility of SCT as a method of assessment for the future. Participants provided responses to each

question on a 5 point Likert scale with 1 labeled “Not at all” and 5 labeled “Extremely.” Participants were also invited to provide written comments on any aspect of the PSCT experience.

At the completion of each site administration, participant’s electronic PSCT response files were saved and transferred into the study database at the home research site.

Statistical Analysis

Each resident’s PSCT was electronically scored using the scoring key established by the expert panel of reference. Raw scores were subsequently transformed as described by Charlin et al.[17] A one-way analysis of variance (ANOVA) was used to determine if the panel of experts obtained higher PSCT scores compared to senior (PGY 3-4) pediatric residents and if senior (PGY 3-4) pediatric residents obtained higher scores than junior (PGY 1-2) pediatric residents. Results were deemed to be statistically significant at the 0.05 level. Effect sizes were calculated using Cohen’s d. The reliability of the PSCT scores was calculated using Cronbach’s α coefficients. To accommodate classical and evolving style case presentations as well as variability in the number of questions per case, questions were used as the unit of measurement. Reliability results were compared to the minimum “qualifying examination standard” of 0.80.

Participants’ responses to the post-test survey questions were reported using Likert scale frequencies (Q1-Q5) and percentages based on case type preference (Q6). Qualitative responses were analyzed by two of the investigators using thematic analysis.[22] The most frequent themes emerging were identified. Representative quotes for each theme were selected and reported.

2.4 Results

Participants

A total of 112 participants completed the PSCT. Twenty-one in-patient pediatricians (of a possible 26) served on the panel of experts (response rate 81%) including 16 women and 5 men. Members of the expert panel had served as staff for an average of 8 years (range 3-22 years). Ninety-one pediatric residents (of a possible 130) participated in the study (response rate 70%) including 72 women and 19 men. The number of resident participants was: University of Calgary (n = 40), University of British Columbia (n = 26), University of Alberta (n = 15) and University of Saskatchewan (n = 10). The number of pediatric residents by post-graduate year (PGY) was: PGY-1 (n = 33), PGY-2 (n = 17), PGY-3 (n = 23), PGY- 4 (n = 18), and by functional training level was: pediatric junior residents (PGY-1 and PGY-2; n = 50) and pediatric senior residents (PGY-3 and PGY-4; n = 41).

Time to Completion

All members of the expert panel completed the PSCT in 90 minutes or less. The range was 57 - 90 minutes. Seventy-seven (77) residents (85%) completed the test in 90 minutes or less. Fourteen residents (15%) required extra time: 8 PGY-1s, 4 PGY-2s, 1 PGY-3 and 1 PGY-4. The residents displayed a wide range of completion times: 42 - 121 minutes. For the purpose of standardized scoring, all responses received by the 90-minute mark were used to calculate each participant's final PSCT score.

Score Analysis: Inclusion/ Exclusion

The final analysis included a total of 12,163 resident responses and 2,877 panel of expert responses. A total of 304 responses (2.0%) were excluded from the analysis as these were received after the PSCT target time of 90 minutes.

Scores

The mean PSCT score of the panel of experts was 80.00 (Range: 68.03 - 87.83; SD = 5.00). The mean PSCT score of senior residents was 70.90 (Range: 54.42 - 80.41; SD = 6.73). The mean PSCT score of junior residents was 65.14 (Range: 29.67 - 80.41; SD = 10.69).

A one-way ANOVA was performed on the three sets of PSCT scores for the panel of experts, senior residents and junior residents. Overall, there was a difference in performance across levels of training, $F = 22.84$ (df = 2); $p < 0.001$. The panel of experts scored higher than both the senior residents (mean difference 9.15; $p < 0.001$, Cohen's d = 1.54, r = 0.61) and the junior residents (mean difference = 14.90; $p < 0.001$, Cohen's d = 4.03, r = 0.90) and the senior residents scored higher than junior residents (mean difference = 5.76; $p < 0.002$, Cohen's d = 1.18, r = 0.51). When sub-divided by single post-graduate years there were no significant differences between the PGY-1s and PGY-2s or between PGY-3s and PGY-4s. The reliability of the PSCT scores (Cronbach's α coefficient) was 0.85.

In addition to the study test administrations, three hypothetical PSCTs were performed to explore if a candidate providing only extreme responses (at each end of the Likert scale) or only neutral responses (middle of the scale) could increase their PSCT scores. In all cases the resulting PSCT scores were less than 35, representing scores far below the average scores of any of the study groups.

Satisfaction Survey Responses

All 112 participants completed the post-test satisfaction survey. The following questions were asked: Q1: “Do you believe this SCT depicts "real-life" clinical decision-making?” Q2: “Do you think this SCT fairly represented the domain of pediatric acute care medicine?” Q3: “Do you like SCT as a new method of measurement?” Q4: “Do you think SCT cases covered a range of difficulty?” Q5: “Would you find it useful to utilize this SCT method of assessment in the future?” (Q1-Q5 results are displayed in Figure 3). Participants were asked a final question in the survey, Q6: “With regards to the classical case type vs. the evolving case type, did you prefer one type of case over another?” (See Figure 4).

Participants also provided a variety of qualitative comments. Primary positive themes included the realism of the cases, the web-based format, the integration of multi-media accessories and the attraction to a test that assesses real and relevant clinical decision-making. Primary negative themes included the potential for multiple interpretations of cases or questions, the challenges in using a 5-point Likert scale, adjusting to different scales for diagnosis, investigation, and treatment (within the same test) and the inability to “go back” to view a previous response in the test (technical limitation). Five comments of particular interest were as follows: 1) “Integration of technology (audio, video and images) made it far more realistic.” 2) “There should be a place for junior trainees to put an ‘I don't know’ option.” 3) “I found evolving cases better simulated real clinical practice.” 4) “I wish we did this type of thing regularly through-out residency and received feedback to gauge our progress and problem areas” 5) “Perhaps if this testing method was used in conjunction with more traditional multiple choice and short answer you could get an overall better representation of knowledge and clinical acumen.”

2.5 Discussion

The main finding of the study is that the PSCT scores were able to discriminate clinical reasoning ability between staff and two distinct training levels (senior and junior residents) in pediatric residency, thereby supporting the construct validity of the PSCT scores. This finding is consistent with other studies in the SCT literature involving staff and different levels of residents in other fields of medicine.[8,9,16,23-29] In contrast to the recent SCT "test response validity" concerns of Lineberry et al, [11] the three hypothetical tests we conducted, to explore the possibility that blindly selecting only 'extreme or neutral responses' could increase test scores, revealed exceedingly low scores, demonstrating that the design of this PSCT was robust enough to protect from such threats. Validity of the PSCT scores was also strongly supported by the results of the post-test survey. Participants, including inexperienced trainees through to senior level staff reported that this PSCT covered a range of difficulty, fairly represented the domain of pediatric acute care medicine and accurately depicted "real-life" clinical decision-making. They also expressed a strong positive response to the PSCT experience and believed it would be useful to utilize the SCT method of assessment in the future. Validity of the PSCT scores was further enhanced by the fact that, independent of the study site, the PSCT was able to discriminate across residents within different pediatric residency programs and geographic locations. These results support the representativeness of PSCT participants and suggest the PSCT is generalizable to other pediatric residency programs in Canada.

Our study also revealed that it is possible to achieve a high level of reliability (Cronbach's α = 0.85) on a PSCT. This may be facilitated when specific strategies in test design and development are applied. In this study, three general pediatricians possessing clinical experience in acute care pediatrics as well as knowledge and experience in test development (including writing SCT cases)

contributed to the PSCT design. Test developers adhered to established SCT construction guidelines. A test blueprint and a series of cases and questions were developed based on acute care topics derived from national pediatric sub-specialty training objectives. Cases were balanced to reflect the type of clinical decisions a pediatrician must make (diagnosis, investigation and treatment) and the flow of medical decision-making (introduction of the evolving case). We aimed to develop an optimal combination of cases (20-24) and questions (3-5) and a range of difficulty (across the training spectrum). We intentionally recruited robust sample sizes for the expert panel and the pediatric resident population at all training levels. Members of the expert panel were required to have met two significant and distinct standards. First, all were required to be certified as specialists in Pediatrics by the Canadian Royal College of Physicians and Surgeons of Canada. Second, all panel members required a minimum of three years of clinical experience as in-patient staff. Is it proposed that these two distinct requirements aided in creating increased "distance" from the senior resident group, and therefore, sensitivity in detecting potential differences. We recruited residents from nationally accredited pediatric specialty training programs. We utilized the University of Montreal web-based SCT design to standardize administration of the test to all participants. Finally, we applied consistent standards for SCT scoring including application of the aggregate scoring method and score transformation.

One of the most interesting and potentially important observations of this study was that while the panel of experts and vast majority of senior residents were able to complete the PSCT within the targeted 90-minute time frame, a significant proportion of junior residents struggled with this task. Could it be that speed helps differentiate clinical reasoning ability? Does increasing clinical experience allow not only more effective clinical decision-making but also more efficient clinical decision-making?

Is the time required to make clinical decisions relevant and important in an acute care environment? If so, at what stage(s) of training should the assessment of clinical decision-making efficiency occur? One theory that would link and support the "experience – efficiency association" is 'script theory' which would imply that by virtue of clinical experience, members of the expert panel and the more senior trainees have developed prototype scripts as well as accumulated an extensive series of exemplar scripts.[30] Armed with this rich framework, they are able to draw more readily on this applied knowledge and efficiently select the most salient features. In contrast, junior residents are still learning to recognize and apply basic prototype scripts. These junior trainees lack the exposure and associated volume that comes with increasing clinical experience, have a fewer number of exemplars to draw on (both typical and atypical cases), and, therefore, may take longer to reason and be less adept at making clinical decisions - especially in situations where there is ambiguity or missing information. While speed of clinical reasoning and decision-making may be less relevant in some medical specialty contexts, one can argue that in acute care situations and especially in urgent or emergent scenarios, this skill is highly relevant and necessary to achieve successful patient outcomes. This skill requires the ability to quickly discern the most salient clues, identify missing information (and order relevant timely investigations), integrate new incoming information (from the patient and the results) and at the same time initiate and gradually focus patient treatment. Given that competency in this skill is required by the time one becomes an independent practicing acute care physician, we propose the assessment of clinical reasoning efficiency in urgent and less complex scenarios occur at the junior resident stage and assessment of clinical reasoning in emergent and more complex scenarios occur at the senior resident stage of training.

Some might suggest that variables independent of clinical reasoning ability may have influenced PSCT test-taking speed such as the lack of familiarity or experience in taking a computer based test, time needed to adjust to the SCT format and skills required to navigate between the questions and the multi-media accessories. Countering these possibilities is that all participants in this study (including the panel of experts and the residents) were naïve to the SCT format – none had ever taken an SCT before. All participants also received practice cases and questions prior to the PSCT. Therefore, if one or more of these variables had been active, one would have expected all participants to be equally affected. As participants gain further experience with SCT, the potential influence of methodological "learning factors" should dissipate.

Another new insight offered by this study is the potential value of the evolving style case. We propose this style is closer to the true sequential timing and flow of medical practice. A second perceived advantage of the evolving case style is that the impact of the single new piece of information is evaluated uniformly across the original set of differential diagnoses. This feature also more closely simulates clinical reasoning and applications to clinical decision-making in real life. Not surprisingly, of the participants who expressed a case style preference (50%), all panel of expert members and a significant majority of residents preferred the evolving style case. This result begs the question … should SCT developers consider increased utilization of the evolving style case in future testing? We believe the answer is "yes".

Finally, the post-test survey results offered some constructive comments for future SCT design. To begin, it is important to ensure that cases are worded and questions are framed such that only one interpretation is possible. This can be challenging but is vital to ensure both the expert panel and residents understand exactly what is being asked. We suggest two strategies be considered. First, test

developers should be highly selective and precise about the information offered and the options provided. Second, it is important to ensure the test is piloted with participants at different levels to identify any discrepancies in interpretation. With respect to the Likert response scales, two changes are proposed: a) use a consistent scale for the entire test regardless of case decision type (diagnosis, investigation, treatment) and b) provide an "I don't know" option. It would also be helpful if respondents could "go back" to review the questions and see (but not change) their responses *within* an individual classical or evolving style SCT case. This also more closely approximates real life where information presented previously is still available to review and not "lost" in the process of clinical reasoning and decision-making. Finally, it may be quite valuable to add a "think aloud" supplement to some cases such that the richness of the respondents' qualitative answers may be explored at the same time the clinical reasoning thought process is being applied.

Limitations

The results of this study should be interpreted in the context of the following characteristics and limitations. First, there were significant variations in our demographic results. Both the panel of experts and resident participant groups contained more women than men (77% and 80%, respectively). However, this high proportion of women is typical of pediatric staff and trainee demographics across Canada. Our study also had a slightly lower proportion of senior residents than junior residents. This is also expected given that some R4s have already left their general pediatrics residency program and chosen to pursue sub-specialty training. The sample sizes in each individual PGY year as well as the combined sample sizes comprising the junior and senior levels were sufficient to provide valid comparisons between these groups. Finally, the number of participants from each site

varied with the larger residency programs (2 sites) contributing substantially more participants. However, no differences in the results were detected between geographic study sites.

A second limitation relates to selection of the members of the panel of reference. A convenience sample of 21 general pediatricians known to work full-time in pediatric in-patient medicine at the home site was used. This sample was not randomly selected and so may have not been representative of general pediatricians working in pediatric in-patient medicine at the other sites. As responses from the panel of experts form the scoring key, there is a risk that how a practice group of pediatricians at one site will score any particular case or question may reflect local approaches, guidelines, styles or biases. To evaluate this possibility, scores of local residents were compared with: a) the scores of residents from each of the other individual sites as well as with b) all other sites combined. Since there were no statistical differences, it can be concluded that the local panel of experts served as a reasonably unbiased and representative panel of experts for this study.

A third limitation is that it was not possible to conduct simultaneous site administrations of the PSCT due to the need for in-person test orientation, website/computer trouble-shooting and variable timing of resident academic half-days. Two weeks were required at the local site to administer the test to the panel of experts, the 4th year residents and the PGY1-3 residents during their academic half-day. Test administration at each of the external sites followed with a single sitting at each site, one week apart. This situation introduced the potential risk of exposed case content or questions. To mitigate this risk, at the end of each test administration expert panel members and residents were specifically asked to maintain strict confidentiality on all aspects of the PSCT. While it does not rule out the possibility that case content was exposed, it is notable that based on the results between

sites, there was no trend of increasing scores for any site or any PGY sub-group with successive test administrations.

2.6 Conclusion

The findings of this PSCT study contribute to a growing body of literature suggesting that the script concordance test holds promise as a valid, reliable and feasible method to assess the core competency of clinical reasoning in medicine. Pediatric staff members and residents also express keen interest and engagement in this form of assessment. Enhancements to SCT may include specific modifications to test design to improve clarity and more fully delineate participant responses, consideration of intentional use of PSCT case load to discriminate clinical reasoning efficiency and increased utilization of the evolving style case. We propose the PSCT may be effectively and efficiently integrated into formative residency assessment and with increasing exposure, experience and refinement may soon be ready to pilot within summative assessments in pediatric medical education.

2.7 Figures

Figure 1: The Classical Style SCT Case – Sample One

A 2-year-old boy presents to the emergency department with a five-day history of fever up to 38.6 degrees Celsius, enlarged cervical lymph nodes, inflamed conjunctiva and a red tongue.

If you were thinking of …	**And then you find** …	**This hypothesis becomes …**
(A diagnostic hypothesis)	(New clinical information Or a test result)	(Select one response)*
Kawasaki's Disease	Echocardiogram report is normal	-2 -1 0 +1 +2
Group A Streptococcus	Swollen and erythematous tonsils with no exudate	-2 -1 0 +1 +2
Mononucleosis (EBV)	Liver palpable at 5cm below the costal margin.	-2 -1 0 +1 +2
Polyarteritis nodosa	Magnetic resonance venography shows mesenteric artery aneurysms	-2 -1 0 +1 +2

*** Scale:** -2 = very unlikely, -1 = unlikely, 0 = neither likely nor unlikely, +1 = more likely, +2 = very likely

Figure 2: The Evolving Style SCT Case – Sample One

Base Scenario: A seven year-old boy with known Type 1 diabetes presents to the Emergency Department. On history the mother reports he awoke from sleep complaining of feeling sick and having "tummy pain". His temperature at home was 37.6 C. She subsequently brought him to the Emergency Department. Since his arrival to the department 4 hours ago he has vomited twice and had one episode of non-bloody diarrhea. Vital signs are: HR = 132, RR = 20, BP = 88/60. Temperature = 37.9 C. Physical examination reveals a boy who appears ill and moderately dehydrated. He has bowel sounds and intermittent peri-umbilical abdominal pain with some rebound tenderness. The remainder of his physical examination is normal. He is receiving a normal saline bolus of 20cc/kg.

Considering the diagnosis of …	**With the information provided, this hypothesis becomes …**				
Intussusception	-2	-1	0	+1	+2
Gastroenteritis	-2	-1	0	+1	+2
Diabetic ketoacidosis	-2	-1	0	+1	+2
Appendicitis	-2	-1	0	+1	+2

***Case Evolution*:** You order some blood tests, a urine sample and an abdominal ultrasound:

Hemoglobin: 122g/L; WBC: 8.2 x 10^9/L, Neutrophils of 6.3 x 10^9/L and Lymphocytes of 4.8 x 10^9/L.
Platelet count: 228 x 10^9/L. Sodium 137 mmol/L; Potassium 3.9 mmol/L; Chloride 100 mmol/L.
Blood glucose of 11.8 mmol/L.
Presence of 2+ ketones in the urine. Abdominal ultrasound report is normal; appendix was not visualized.

Considering the diagnosis of …	**With the information provided, this hypothesis becomes …**				
Intussusception	-2	-1	0	+1	+2
Gastroenteritis	-2	-1	0	+1	+2
Diabetic ketoacidosis	-2	-1	0	+1	+2
Appendicitis	-2	-1	0	+1	+2

***Scale**: -2 = very unlikely, -1 = unlikely, 0 = neither likely nor unlikely, +1 = more likely, +2 = very likely

Figure 3: The PSCT Post-Test Survey

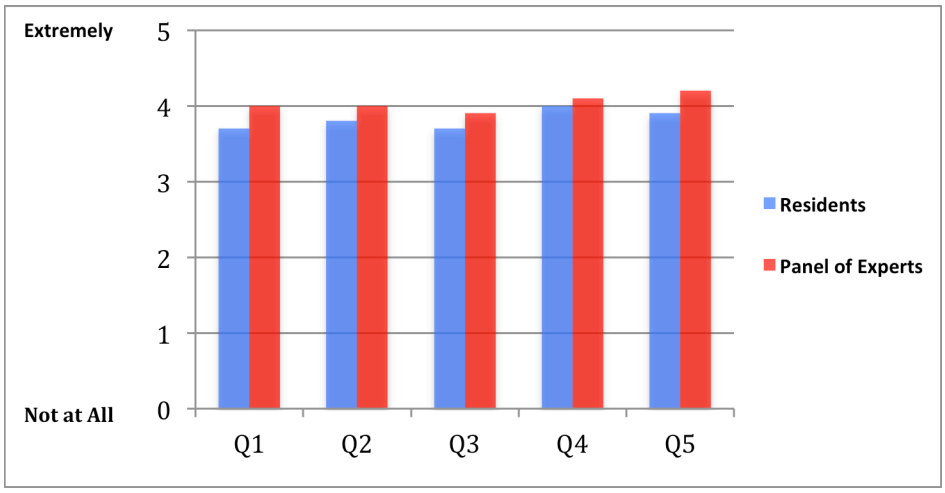
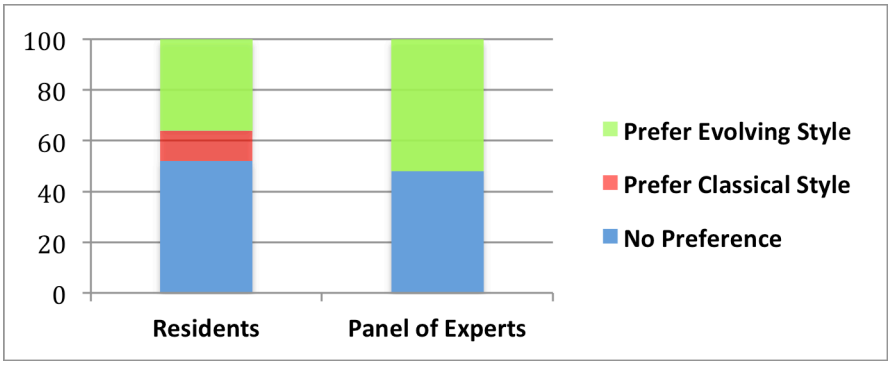


Figure 4: Preference for Case Style: Classical vs. Evolving



Chapter Three

Script Concordance Testing and the Evolving Case Type – A Closer Look

The results of the overarching PSCT study provided concrete answers to three pre-determined questions. This study revealed that PSCT scores could discriminate clinical reasoning ability between staff and two distinct training levels in pediatric residency. This study also demonstrated that it is possible to achieve a high level of reliability in scores, particularly when specific strategies that target test design, development and administration are applied. The third major finding was that staff and trainees found this type of assessment method very engaging, express interest in its utilization in the future and have offered constructive suggestions for improvement.

One of the less anticipated and yet most intriguing results of the overall study was the participants' response to the introduction of the *evolving* SCT case design. Of the PSCT participants who expressed a case type preference (50%), all members of the expert panel and a significant majority of residents preferred the evolving type case. This was a surprise. This trend was reinforced by several qualitative comments offered in the post-test survey with one of the emerging themes being related to the realism associated with the evolving case design. Reflecting on these results, we recognized the significant potential for increased utilization of the evolving type case in the future. This led to the impetus to delve further into the data to conduct a more in-depth analysis of the quantitative and qualitative results from the evolving case type data. A number of interesting questions were raised:

1) Would the scores of the PSCT evolving case data independently be able to discriminate clinical reasoning ability between staff and two functional levels of pediatric residents?
2) Would the scores of the PSCT evolving case data produce robust reliabilities?
3) What specific themes emerged from comments made by participants about the evolving type case?
4) What were the original reasons for proposing the evolving type case design?
5) During the process of creating and reviewing the results of the evolving type cases, what new insights were acquired?
6) What can SCT developers do to improve the design and construction of the evolving type case in the future?

Finally it was believed that if the evolving case design were to be incorporated into future SCTs, test developers would need a more detailed explanation of the principles and the process of how to construct this case type.

In light of the above questions and goals, the primary purpose of the second paper was to conduct a focused analysis of the evolving type case data and attempt to answer the first three questions outlined above. The secondary purpose was to state the original underpinnings that led to introducing the evolving type case in this study and to explore if these and/ or other theories might explain the results that were obtained. The final purpose was to explain the similarities and differences between the classical and evolving style case designs with the goal of better understanding the evolving type case and, in essence, provide an initial primer so that other SCT developers could begin to create evolving type cases for their own field of clinical medicine.

Chapter Four

Script Concordance Testing and the Evolving Type Case:

Is There a New Kid on the Block in Clinical Reasoning Assessment?

4.1 Abstract

Context: Script concordance testing (SCT) is gaining momentum as a viable method of assessment of clinical reasoning. We introduced a new type of SCT case design, the evolving style case. We proposed that the evolving case style may more closely represent the true sequential timing and flow of medical practice, and therefore more closely simulate the reasoning processes of clinicians.

Objectives: We aimed to: 1) determine whether the evolving style SCT case could help differentiate clinical reasoning ability among practicing physicians and two functional levels of trainees, 2) evaluate the reliability scores of SCT evolving style cases and 3) obtain feedback from participants to help inform the acceptability of the evolving style case in SCT assessment.

Methods: A 24 case SCT was administered to 50 junior residents (JR) and 41 senior residents (SR) from four residency training centers. Twenty-one pediatricians served on the panel of experts (POE). A sub-analysis across the three levels of experience was conducted on the 12 evolving style cases. A post-test survey was used to obtain participant feedback. Responses were analyzed using thematic analysis.

Results: A difference in performance existed across levels of training, $F = 19.31$ ($df = 2$); $p < 0.001$. The POE scored higher than the SR (mean difference = 10.34; $p < 0.001$) and the JR (mean difference = 16.00; $p < 0.001$). SR scored higher than the JR (mean difference = 5.66; $p < 0.001$). The reliability (Cronbach's α) of the evolving style case scores was 0.83. The most prevalent theme was that participants find the evolving case style more true to real clinical decision-making.

Conclusions: Our findings suggest that the evolving style case can be highly effective in distinguishing clinical reasoning ability across three levels of experience. Participants find the evolving style case engaging and perceive the clinical reasoning process required in this case design to be more similar to the decision-making processes of clinical practice. We suggest increased utilization and refinement of the evolving style case may help to support SCT as an increasingly robust, engaging and relevant method for the assessment of clinical reasoning in medicine.

4.2 Introduction

Script concordance testing (SCT) is an emerging method within the domain of clinical reasoning assessment in medicine.[6,17,28,31] SCT is intentionally designed to assess a candidate's ability to reason when faced with challenging decisions typically encountered in the three main phases of clinical decision-making: diagnosis, investigation and treatment.[7] The SCT method is relatively unique in the domain of clinical reasoning assessment as it combines two important and valuable features: 1) SCT is designed in the context of authentic clinical scenarios; the realism of the assessment experience is further enhanced by high fidelity accessories including video, audio and clinical images. 2) In contrast to other labor-intensive methods of assessment that target higher-order thinking, SCT is highly efficient and feasible. The assessment consists of a 60 - 90 minute web-based test that can accommodate multiple candidates at once, is low-cost, requires minimal supervision and provides instant results.[21]

Traditionally, SCT has followed a format whereby a clinical case scenario is presented followed by a series of questions: "if you were thinking "x" and then you learn "y", the likelihood of the impact on a clinical decision is "z".[9] We have called this the "classical" SCT case style (see Figure 1). We created and introduced a new type of SCT case style called the "evolving" type case that follows a

two-step approach (see Figure 2). For the purposes of this explanation, a "diagnostic phase" scheme is used. In the first step an initial case scenario and a series of diagnoses are presented. Based on the information in the case scenario (only), the participant is asked about the likelihood of each diagnosis. In the second step, new, subsequent information about the same case scenario is presented (an update on patient's clinical course, results of an investigation or a response to treatment). The participant is then asked to evaluate the impact this new information has on the original set of diagnoses or a subsequent clinical decision regarding the same case.

Differences in Constructing the Evolving Style Case

When designing either the classical or evolving SCT case types, the same basic principles described by Charlin et al.[21] are followed: a) the test is case based, b) short clinical scenarios are created that contain uncertainty, c) sufficient information is provided to generate varying degrees of plausibility and acceptability of response options, and d) the opinion of the panel of experts forms the scoring key. The first three steps of construction are also common: 1) determine the purpose of the evaluation and define the assessment objectives, 2) identify clinical situations that adapt to the assessment objectives, and 3) choose clinical situations that are representative of reasoning challenges typically experienced within that discipline. The situation must reproduce an "authentic clinical context" that is problematic, even for an expert. Thereafter the classical and evolving case styles diverge. In the ***classical case***, the scenario is presented and 3 to 4 different ("if you were thinking") hypotheses are stated. After each of these hypotheses, new clinical information is provided (and is different for each item). A judgment of the impact of the new information on the likelihood of each hypothesis is then made – essentially 3 to 4 different questions per case. A five point Likert scale

is utilized ranging from -2 (Very Unlikely) to 0 (Neither Less Likely or More Likely) to +2 (Very Likely). In the ***evolving style case*** there are 2 stages. In the first stage, the scenario is presented, no new information is offered and the likelihood of each of 4-5 hypotheses is rated based on information provided in the initial scenario. In the second stage new clinical information is provided (a change to the patients clinical presentation, a lab result or the response to a treatment) and this information is applied to the original set of diagnostic hypotheses, or, to a set of relevant hypotheses for a subsequent realistic clinical decision (investigative or treatment) for the same case. The new piece of information offered must be clear, specific and impact the likelihood of the clinical options provided. During this step the participant must integrate the original scenario and the new information, analyze the data within the updated context and evaluate how the new information influences their judgment about possible diagnoses, investigations or treatments.

In summary, the goal of the first stage of the evolving case is to assess the participant's ability to identify, classify and interpret clinical data and to apply it within a specific context. The goal of the second stage is to evaluate how one new piece of information impacts clinical reasoning applied over either the original series of options (reconsideration of a clinical decision in light of new information) or how the new information globally impacts hypotheses associated with a subsequent stage of clinical decision-making.

We proposed that the evolving style case may more closely represent the true sequential timing and flow of medical practice, and, therefore, might more closely simulate the reasoning processes of clinicians. We sought to: 1) determine whether the PSCT evolving style cases could help differentiate clinical reasoning ability among practicing physicians and trainees, 2) evaluate the reliability of scores

of the PSCT evolving style cases and 3) obtain qualitative feedback from participants (clinical staff and trainees) to help inform the potential acceptability of the evolving style case in SCT assessment.

4.3 Methods

Three Canadian Royal College certified general pediatricians constructed the Pediatric Script Concordance Test (PSCT). These pediatricians possessed a minimum of seven years of clinical in-patient experience. All had formal training and experience in trainee test development and were familiar with SCT format and methodology. The guidelines for construction of the PSCT followed those described by Fournier et al.[19] The Royal College of Physicians and Surgeons of Canada (RCPSC) Pediatrics "Objectives of Training" served as the basis for the test blueprint.[20] This blueprint was applied across 24 SCT cases: 12 classical - style and 12 evolving - style. During the development of cases and questions, specific efforts were undertaken to: a) ensure a wide variety of cases typically seen in pediatric in-patient medicine, b) address the three primary clinical decision-making situations: diagnosis, investigation and treatment, c) embed varying levels of uncertainty to realistically represent clinical decision-making, and d) appropriately challenge trainees across all four years of the residency training program.

Approval for this study was sought and obtained from the research ethics boards at each of the four respective university study sites. The University of Montreal SCT web-based design was utilized to administer the PSCT.[21] Integration of audio (heart sounds), visual images (x-rays, rashes, a growth chart and an ECG) and video (a child with respiratory distress and an infant with abnormal movements) was achieved via a pre-loaded USB stick. Within selected cases of the PSCT, participants were prompted to access these relevant accessories.

Resident responses to each question were compared with the aggregate responses of the panel of experts. Questions had the same maximum (1) and minimum (0) values. All questions of the PSCT were equally weighted. A score of 100% would mean that the participant answered each question exactly like that of the majority of members from the expert panel. Following the recommendations by Charlin et al, final scores were calculated by taking the raw scores of all participants and transforming these into *z* scores and *T* scores. Calculations were performed using the standard deviation of the panel of reference.[17]

Participants

The panel of experts (POE) was recruited from a group of in-patient general pediatricians from the local study site. All had been certified by the RCPSC and possessed a minimum of three years of full-time clinical experience in general pediatric in-patient medicine. Pediatric residents from four universities in Western Canada were recruited to participate in the study. Residents from all post-graduate years (1-4) were included in the recruitment. The primary investigator introduced the research study to the panel of experts during a monthly staff meeting and to local residents during their weekly academic half-day. The study was introduced to residents at 2 additional sites by video teleconference and slide presentation. A local presenter was trained to deliver the same informational session at the fourth site. All groups received a 30-minute orientation to the SCT format and to the classical and evolving style cases. An email invitation was sent to each potential participant following the informational sessions. Participants provided written consent prior to test administration.

PSCT Pilot and Optimization

The PSCT was piloted to assess: a) test content and duration and b) technical feasibility. Participants completed a post-test survey in which they were asked to comment on test readability, perceived interpretation of cases and questions, and, perceived difficulty. Pilot test duration times were also recorded. Technical feasibility included: a) maintenance of the Internet connection to the web-based site and, b) perceived ease of navigation between USB accessories and the PSCT web cases. The information obtained from the pilot served as the basis for optimization of PSCT cases and questions. The PSCT pilot version consisted of 31 cases (16 classical and 15 evolving) and 186 questions. A total of 7 cases and 49 questions were removed for the following reasons: two cases were found to have multiple interpretations, two cases were deemed to be excessively long or complex, one case was judged too easy and two cases were removed to reduce test length. The optimized (final) version of the PSCT consisted of 24 cases (12 classical and 12 evolving) and 137 questions. The evolving cases accounted for 95 questions.

PSCT Administration

During a two-week period in February, 2013 the PSCT was administered to the panel of experts. Over the following three-week period the PSCT was administered to pediatric residents during their academic half-day at each of the four western Canadian university sites. The co-investigator and a research assistant provided oversight to all test administrations. Each testing session began with a 20-minute orientation including: 1) an orientation to the session (welcome and thank-you to participants, introduction of research personnel, session agenda), 2) a summary of the SCT concept and on-line testing format, 3) a review of the classical and evolving type cases, 4) a reminder about the test scope (acute care in-patient general pediatrics), test scale (number of cases and questions) and target test

time (90 minutes), and, 5) instructions for navigation between the PSCT website and the USB stick. Each participant logged into the PSCT website and proceeded to complete the test. The web-based program tracked individual responses and time. The test administrators also tracked PSCT completion times. Final scores were calculated based on the responses received by the 90-minute mark. For the purposes of this analysis, scores were derived from responses to the evolving cases only.

The PSCT was followed by a 10-minute web-based post-test survey. This survey included a question specifically designed to address the classical case vs. the evolving case designs: "With regards to the *classical case type* vs. the *evolving case type*, did you prefer one type of case over another?" Participants were also invited to provide qualitative comments on any aspect of the PSCT experience. Upon completion of each test site administration, participant's response files were saved and transferred into the study database at the home research site.

Statistical Analysis

PSCT Scores

Each resident's PSCT was electronically scored using the scoring key established by the expert panel of reference. One-way analysis of variance (ANOVA) was used to determine: a) if the panel of experts obtained higher PSCT ***evolving*** case scores compared to senior (PGY 3-4) and junior (PGY1-2) pediatric residents, and b) if senior pediatric residents obtained higher ***evolving*** case scores than junior pediatric residents. Results were deemed to be statistically significant at the 0.05 level. Effect sizes were calculated using Cohen's d. Reliability of the PSCT ***evolving*** case scores was calculated using Cronbach's α coefficients. Reliability results were compared to the minimum "qualifying examination standard" of 0.80.

PSCT Survey Responses

Participants' responses to the post-test survey question pertaining to PSCT case types were analyzed using frequencies. All qualitative comments pertaining to the evolving case types were coded for primary themes using thematic analysis.[22] Representative quotes for each primary theme were identified.

4.4 Results

Participants

Ninety-one residents and 21 experts completed the PSCT. Response rates were 70% (n= 130) and 81% (n= 26), respectively. Members of the expert panel had served as full-time staff for an average of eight years (range 3-22 years). The number of pediatric resident participants was: University of Calgary (n = 40), University of British Columbia (n = 26), University of Alberta (15) and University of Saskatchewan (n = 10). The number of pediatric residents by post-graduate year (PGY) was: PGY-1 (n = 33), PGY-2 (n = 17), PGY-3 (n = 23), PGY- 4 (n = 18), and by functional training level was: pediatric junior residents (n = 50) and pediatric senior residents (n = 41).

Scores

Each participant answered all of the evolving case type questions (there were no missing data). Fourteen residents required extra time to complete the PSCT. The maximum test core was 100. The evolving case type score mean of the panel of experts was 80.00 (Range: 69.14 – 89.19; SD = 5.00), senior residents 69.66 (Range: 50.34 - 80.41; SD = 7.65) and junior residents was 64.00 (Range: 21.52 - 77.91; SD = 12.70). A one-way ANOVA was performed on the three sets of evolving case type scores

for the panel of experts, senior residents and junior residents. Overall there was a difference in performance across levels of training, $F = 19.31$ (df = 2); $p < 0.001$. The panel of experts scored higher than both the senior residents (mean difference = 10.34; $p < 0.001$, Cohen's $d = 1.60$, $r = 0.63$) and the junior residents (mean difference = 16.00; $p < 0.001$, Cohen's $d = 1.66$, $r = 0.64$). The senior residents scored higher than the junior residents (mean difference = 5.66; $p < 0.001$, Cohen's $d = 0.54$, $r = 0.26$). There were no significant differences among the individual year cohorts (PGY1-2) and (PGY3-4). The reliability (Cronbach's α coefficient) of the evolving case type score was 0.83.

Survey Responses

All 112 participants completed the post-test survey. All responded to the specific question about the evolving case type: "With regards to the classical case type vs. the evolving case type, did you prefer one type of case over another?" For the residents, 52% (n = 47) had no preference, 36% (n = 33) preferred the evolving cases and 12% (n = 11) preferred the classical cases. For the panel of experts, 48% (n = 10) had no preference, 52% (n = 11) preferred the evolving cases and none preferred the classical cases.

Participants offered several qualitative comments about the evolving type case. The most prevalent theme was: 1) "I found evolving cases more true to real clinical decision-making." Other comments included: 2) "Evolving cases were useful for ranking various options – which is closer to the thought process we often use – especially to decide on a "working diagnosis." 3) "Evolving cases were easier to follow than the classical cases." 4) "I really liked the evolving cases - each case is like a patient's story." 5) "For the evolving questions especially, it would be nice to be able to see the options in each section all at once instead of one at a time. Also – I would like to be able to view my original responses from the first section when I am completing the second section. That chance to

"compare" would be more similar to how we do it in practice." 6) "For both the classical and evolving cases I would like to be able to have a comment box where I can explain what I am thinking to qualify my answer." 7) "For the evolving cases it is sometimes difficult to determine whether I am answering based on the new information only or based on the whole case."

4.5 Discussion

Results from this research contribute to the field of SCT and clinical reasoning assessment in two new and important ways. First, we discovered that the evolving type SCT case could help differentiate clinical reasoning ability between practicing physicians and two functional levels of resident trainees. Second, we learned that participants found the evolving type case engaging, easier to follow, closer to the reality of decision-making in clinical practice and, overall, either equivalent or preferable to the classical SCT case design. Each of these findings will be discussed separately.

Differentiating Clinical Reasoning Ability

Multiple studies have demonstrated that a classical case-based SCT exam can differentiate clinical reasoning ability between practicing physicians and trainees. Based on the results of this study, it appears that the evolving case type can also be utilized to distinguish between these groups. Features of the evolving case method that may enhance its validity is the concept of "flow" and the reality that in clinical medicine, the patient's clinical story evolves "over time". It is useful to assess a trainee's ability to follow the patient's clinical course and examine how their thinking changes as time progresses and new information becomes available. The evolving type case design mirrors this sequential process. Another characteristic that may help support the evolving case as a valid measure of clinical reasoning is the fact that clinical decisions are made based on the "whole" clinical picture at

any given point in time. In the first stage of the evolving case, a series of clinical facts and observations are provided. Clinicians select which features are most relevant and begin to connect these, similar to starting a puzzle. The most recognizable and relevant pieces are selected first; other less recognizable pieces may be held in "standby" mode and still others may be discarded altogether if perceived to be irrelevant. Some pieces may have a more central role and a greater influence on the "overall picture" while other pieces or sections are more peripheral and supplementary. At any given point in time, a clinician must formulate, as best they can, the "big picture" at that stage and make judgments based on that holistic impression. In the evolving case these judgments occur at the end of the first stage. In the *second* stage, the patient's story is updated, the new information (like a set of new puzzle pieces) are applied to the initial impression and may lend even further weight to the initial picture, may alter it slightly or may change it altogether. In essence, the patient's clinical story is "evolving" and with thoughtful integration of new information at each stage, the picture (ideally) becomes increasingly clear. This is the hallmark of the evolving style SCT case. Participants are assessed at two natural stages of clinical decision-making – during an initial assessment and at some point in follow-up. At each stage the clinical picture is viewed in its entirety. During the second stage there is a particularly valuable opportunity to "zero in" and assess whether the candidate is able to integrate a particularly important clinical concept that may significantly impact the patients working diagnosis, next set of investigations or treatment. The ability to integrate and recognize such critical information is a central feature of clinical reasoning and may be why the evolving case may be well suited for the assessment of this core medical expert competency.

A third factor that may enhance the validity of the evolving type case is that it may be more effective at eliciting (and evaluating) higher levels of cognitive functioning, including clinical

reasoning. Consider Bloom's taxonomy of cognitive functioning (see Figure 3) in relation to the evolving case.[32] Each SCT case is created to include typical information that would be available for any given clinical case. For example the respiratory rate is provided as part of the patient's "vital signs". The candidate must recognize this vital sign (knowledge) and interpret it as being slow, normal or fast (comprehension). This rapid respiratory rate (tachypnea) may be presented in the context of a patient with asthma. The candidate must apply this information to this specific context (application), recognizing that tachypnea may be present in many different case presentations, including asthma, pneumonia and cystic fibrosis. Additional information such as the presence or absence of fever, past medical history, a physical examination and a chest x-ray may help to differentiate these potential diagnoses (analysis). In contrast to tests of multiple choice (MCQ) or short answer (SAQ) where knowledge and comprehension are typically assessed, SCT requires the candidate to *integrate* new information and then *apply* and *analyze* it in new contexts. This requires higher levels of cognitive functioning and the skills of clinical reasoning. Furthermore the evolving case helps to assess these clinical reasoning skills within the *context* of genuine medical decision-making by re-visiting a working diagnosis (a common clinical occurrence) or applying and analyzing this information within the context of a subsequent clinical decision – usually related to investigation or treatment. Therefore, by virtue of the potential for the evolving case design to: a) tap into higher levels of cognitive functioning as well as, b) represent authentic clinical decision-making contexts, the evolving type SCT case may better target the assessment of clinical reasoning.

To summarize, the evolving type case is designed to reflect the *evolution* of a patient's course over *time* and to mirror the pattern of *integrating, applying and analyzing* that information *as a whole* at typical *contextual* stages of clinical decision-making. It is for these reasons that the evolving type

case may be “closer to reality” and more representative of true clinical reasoning. These theories may at least partially explain why in this study, the evolving type case was effective in delineating between medical experts and two functional levels of resident trainees in the fundamental skill of clinical reasoning.

The Evolving Type Case and Feedback From the Participants

Although overall comparative scores for classical and evolving type cases were very similar (within 2% for respective groups), the majority of participants expressed either an indifference to case type or a preference for evolving style cases. Only a very small minority of residents (12%) expressed preference for the classical case design. Notably, all members of the expert panel expressed either a preference for the evolving type case or no particular preference. None of the twenty-one member expert panel selected a preference for the classical type case. The qualitative comments helped to delineate some of the reasons why many participants preferred the evolving case. Several participants stated that they felt the evolving case more closely represents the thought processes associated with “real clinical decision-making.” Others liked the flow of the evolving cases and found them easier to follow. One participant commented on the “story-like” nature of evolving cases. Perhaps this aligns better with the inevitable ebb and flow of the human condition and the contextual variables that may impact and influence the patient’s course over time.

Participants offered several constructive comments to help improve the evolving type case. First, it would be helpful if the presentation format of the cases and questions could be altered to allow participants to see the entire first step at once, including all questions. When answering the second stage it is also reasonable to permit the participant to not only review the clinical stem (as is

currently the case) but also all previous options and responses. Both of these adjustments could further increase the realism of the evolving type case. Second, it is essential to make clear to participants that in the second stage, the questions are based on the cumulative information offered up to that point in time. It is true that the new information provided may significantly alter the clinical impression but it is also true that it may change it very little. Participants must make that judgment. This reflects actual day-to-day clinical reasoning and decision-making. Finally, it could be very helpful to offer a comment box at the end of each case that would allow the participant to briefly share their rationale for their decision-making. This could be especially useful in a formative sense if using or reviewing SCTs for teaching purposes. We suggest incorporating these recommendations into the evolving case design of future SCTs. Given that this study represented a "first attempt" at the evolving case design, it is anticipated that for all the reasons proposed, this case design may become more widely used with additional refinement and future experience.

Limitations

The scores in the limited time evolving case analysis are impacted by an order effect; participants completed the 12 classical cases first followed by the 12 evolving cases. Therefore, the scores of the evolving cases (compared with the classical cases) are preferentially affected by the time limit. However, all participants were equally affected by this limitation, and, therefore, scores between groups may be legitimately compared.

A second limitation exists with respect to the naivety of participants to SCT testing and the evolving case type. None of the participants (residents or expert panel members) had ever taken an SCT before, and none had ever been exposed to the evolving case type. While all participants received

identical orientations prior to this SCT, one should interpret the results with some caution. A "learning effect" is anticipated in at least 2 aspects of test taking. One aspect relates to becoming accustomed to the SCT case description and in particular the evolving case pattern. A second aspect relates to familiarization with the 3 different response scales. The third aspect pertains to efficiency. Participants will learn and develop confidence with this type of testing (and the evolving case) at different rates. These aspects may be (at least to some degree) independent of the participant's ability to clinically reason. The advantage of this study is that all participants were equally naïve to SCT and the evolving type case so results may be rightfully compared. However, it would be ideal for participants to obtain more experience with SCT testing and the evolving type case such that any learning effect would dissipate and the differences in scores may be more confidently attributed to differences in clinical reasoning.

4.6 Conclusions

Script concordance testing is gaining momentum as a valid method of measurement of clinical reasoning. We introduced and tested a new type of SCT case design, the evolving type case. Our data suggest that the evolving case can be reasonably effective in distinguishing clinical reasoning ability between two functional levels of trainees and experts. We propose that the validity of the evolving type case may be supported by specific features that enhance realism including evolution of a patient's course over time, intermittent (two stage) holistic assessment, and respect for typical contextual points of clinical decision-making. Our results revealed that both residents and experts find the evolving type case engaging, more like actual patient cases and perceive the clinical reasoning process required in the evolving case design to be more similar to the typical decision-making

processes inherent in true clinical practice. We suggest that additional adjustments to the evolving case design combined with increased utilization, experience and refinement may help to support SCT as an increasingly robust, acceptable and feasible method for the assessment of clinical reasoning in medicine.

4.7 Figures

Figure 5: The Classical Style SCT Case – Sample Two

A 2 year-old child presents to the emergency department with a two-day history of increasing respiratory distress:

If you were thinking of … **becomes ...**	**And then you find …**	**This hypothesis**
(A diagnostic hypothesis)	(New clinical information Or a test result)	(Select one response)*
1-Asthma	Night time cough	-2 -1 0 +1 +2
2-Pneumonia	Crackles and wheeze	-2 -1 0 +1 +2
3-Bronchiolitis	Copious nasal secretions	-2 -1 0 +1 +2
4-Croup	Expiratory wheeze	-2 -1 0 +1 +2

***Scale**: -2 = very unlikely, -1 = unlikely, 0 = neither likely nor unlikely, +1 = more likely, +2 = very likely

Figure 6: The Evolving Style SCT Case – Sample Two

***Base Scenario*:** A 28 year old, G3P2 mother has just delivered a male infant at 39 weeks gestation. The baby was delivered by caesarean section due to failure to progress. Apgar scores were 8 at one minute and 9 at five minutes. Birth weight was 3625 grams. Vital signs taken by the nurse at 30 minutes of life were as follows:

- Heart rate = 146 beats/minute
- Respiratory rate = 58 breaths/minute
- Blood pressure = 50/32 mmHg
- Temperature = 36.8 degrees Celsius
- Oxygen saturation = 89% in room air

The infant is now 1 hour old and attempting to breastfeed. The mother reports that the infant latches and feeds for 30-40 seconds and then stops feeding. During this time he appears to have prolonged pauses in his breathing. These episodes are accompanied by a loss of color in the lips and face.

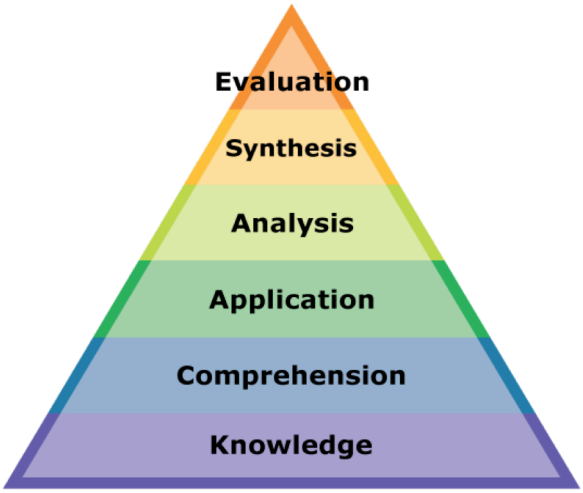
Considering the diagnosis of …	**With the information provided, this hypothesis becomes …**				
Tracheo-esophageal fistula	-2	-1	0	+1	+2
Transient tachypnea of the newborn	-2	-1	0	+1	+2
Congenital heart disease	-2	-1	0	+1	+2
Choanal atresia	-2	-1	0	+1	+2

***Case Evolution*:** You observe the baby feeding for 2 minutes and witness one of these episodes. You notice that the baby's color improves when he is crying.

Considering the diagnosis of …	**With the information provided, this hypothesis becomes …**				
Tracheo-esophageal fistula	-2	-1	0	+1	+2
Transient tachypnea of the newborn	-2	-1	0	+1	+2
Congenital heart disease	-2	-1	0	+1	+2
Choanal atresia	-2	-1	0	+1	+2

***Scale**: -2 = very unlikely, -1 = unlikely, 0 = neither likely nor unlikely, +1 = more likely, +2 = very likely

Figure 7: Blooms Taxonomy of Cognitive Learning [12]



The learning pyramid progresses from the base to the top:

acquisition of basic knowledge (knowledge) ➔
developing a clear understanding of that knowledge (comprehension), ➔
learning how to apply the knowledge in different contexts (application), ➔
being able to analyze the "fit" of information with the clinical case (analysis), ➔
learning how to create new hypotheses and solutions (synthesis) and ➔
developing the skills to be able to effectively evaluate the plan (evaluation).

Chapter Five

Insights Acquired from the Pediatric Script Concordance Test Project:

Implications for Next Steps

To date our goal has been to share the results of the overarching PSCT consisting of classical and evolving style cases. Our primary aim was to evaluate the impact of specific strategies applied to the design of this PSCT and to learn whether or not these measures could influence the validity and reliability of PSCT scores. We also chose to pilot the evolving style SCT case, postulating that the two-step design might more closely represent the true sequential timing and flow of decision making in clinical practice. We were surprised by the potential for this case style to discriminate clinical reasoning ability and the high reliability associated with these test scores. We were also struck by the degree of engagement and preference expressed by expert panel members and residents for this case design. This was impressive considering this was the first time that the evolving case design has ever been used in an SCT. These results provided further motivation to understand why the evolving type case might more closely simulate actual processes of clinical reasoning and decision-making.

In the second paper we sought to understand what potential associations and unifying theories could explain the high degree of resonance displayed by both the quantitative and qualitative evolving case type results. As we began to make some links and postulate some grounded theory between how the evolving case style works and its links to true clinical decision-making, we recognized the opportunity for the clinical practice and assessment "gap" to be addressed. The next natural question became: "What is preventing the integration of this type of assessment into summative testing now?"

While this question was being considered, another highly influential event occurred. The preliminary results of this overall SCT research project were shared via an oral presentation at the 2014 Ottawa Conference, an international conference dedicated to the assessment of competence in medicine and the health care professions. This presentation was part of a series given by researchers engaged in the assessment of clinical reasoning, included other SCT researchers. One of the themes inherent in the SCT presentations was that in SCT there is the potential for two or more answers to be acceptable (and receive credit) because the scoring key is based on the responses provided by the panel of experts. During the post-presentations discussion this theme was challenged by the traditional premise that: "In any high stakes assessment there can only be 'one right answer' to any given exam question." Interestingly, senior exam developers held firmly to the belief; however, practicing clinicians in the audience contested this. During this debate, a clinical example was presented that illustrated a case in which uncertainty prevailed, yet a clinical decision had to be made. The three different (but all reasonable) approaches of three practicing physicians were offered. The senior exam developers categorically stated that this type of example could not be used in a summative evaluation. The practicing physicians argued: "But this is exactly the kind of situation we are faced with in everyday life. If this is the competency required to practice medicine, then why aren't we formally assessing it?" We believe this is a valid question and that if we are to advance the field of clinical reasoning assessment further, we must discover these missing links.

Based on the results of the overarching PSCT, the insights developed from further analysis of the evolving type case and the poignant discourse from the Ottawa Conference on Assessment, we chose, in this third paper to explore three key questions: 1) What traditional philosophical principles of assessment (that are still being adhered to) might be causing barriers to creating greater alignment

between current realities of decision-making and the assessment of clinical reasoning? 2) Can we re-visit these previously held principles in light of contemporary assessment frameworks, recent research and the current gap? 3) Can we formulate strategies to remove these barriers such that formal assessment more closely aligns with the clinical reasoning competencies needed to practice in the realm of clinical medicine in 2015?

Chapter Six

Assessment of Clinical Reasoning in Medical Education: Closing the Gap Between Current Assessment and Reality

6.1 Abstract

Context: Clinical reasoning is a critical skill required to make safe and effective decisions in the practice of clinical medicine. The Royal College of Physician and Surgeons (RCPSC) in Canada and the American Accreditation Council for Graduate Medical Education in the United States (AAC-GME) have requested that methods be developed to assess this core physician competency. Script concordance testing (SCT) is gaining momentum as a viable method of testing clinical reasoning. Yet, if SCT and other potential methods of valid clinical reasoning assessments are to be embraced, some traditionally held beliefs about assessment will need to be reconsidered.

Methods: In this paper we discuss two particular features of SCT that we believe are philosophically paramount to the future assessment of clinical reasoning in medicine: 1) assessment in the context of "uncertainty", and 2) acknowledging that it is entirely possible (and reasonable) to have more than "one approach/ answer" during the course of clinical decision making. We appreciate that attempting to bring these realities of clinical medicine to the field of assessment creates significant challenges. We attempt to identify current barriers and explore potential strategies such that these assessment challenges may be overcome.

Conclusions: We have attempted to explain why uncertainty is an inherent feature of clinical medicine and how it may be entirely valid to support more than one approach to a clinical decision. However, bringing these realities to the field of assessment creates challenges. We propose specific strategies

to help overcome the challenges including: 1) acknowledging situations of uncertainty, 2) creation of clear frameworks that define progressive levels of clinical reasoning skills, 3) collecting validity evidence to increase the defensibility of such assessments, 4) consideration of comparative feasibility of clinical reasoning assessments and, 5) development of schemes to evaluate the impact of these assessment methods on future learning and practice. We believe the engagement of such strategies will help close the gap between current assessment and the requirements associated with effective clinical reasoning and decision-making in medicine.

6.2 Introduction

In an age where physicians have access to an overwhelming volume of clinical information and are faced with increasingly complex and fast paced medical decisions, the ability to discern the most relevant aspects of a case and execute sound clinical reasoning is essential to optimal patient care. If the skill of clinical reasoning is so crucial to modern day physician performance then surely this core competency is: a) systematically evaluated, and b) receives targeted attention in both formative and summative assessment during residency education. But does it?

To date, evaluation is based primarily on the foundational elements of knowledge and comprehension, especially in the context of certifying examinations. Much less attention is placed on the formative or summative assessment of essential, higher order functions associated with the skill of clinical reasoning including the *application* of knowledge and the ability to *integrate* and *analyze* information "in context". As such, a gap exists between what we recognize is required for clinical performance and what is currently being assessed. Recognizing this deficiency, the Royal College of Physicians and Surgeons of Canada (RCPSC)[4] and the American Accreditation Council for Graduate

Medical Education in the United States (AAC-GME)[5] have requested that specific methods be developed to assess the clinical reasoning competency of medical trainees in formative and summative assessments. In pursuit of this effort, we chose to adopt the script concordance test (SCT) method and attempted to create and implement test development strategies that could enhance the validity, reliability, and acceptability of this method.[33,34]

The purpose of this paper is to shed light on two particular features of SCT that we believe are philosophically paramount to the future assessment of clinical reasoning in medicine: 1) assessment in the context of "uncertainty" (when, despite all of the information that is available there is some doubt as to the correct diagnosis, investigation or treatment), and 2) acknowledging that it is entirely possible (and reasonable) to have more than "one correct path / right answer." We recognize that attempting to bring these realities of clinical medicine to the field of *assessment* creates significant challenges. We attempt to identify current barriers and explore potential strategies such that these challenges may be overcome.

6.3 Assessment in the Context of Uncertainty

Traditionally, medical *assessment* has taken place in the context of *certainty*. Trainees are primarily tested on their ability to remember a large number of facts, patterns, associations and algorithms. For example, multiple-choice questions (MCQs) have a single correct answer and short answer questions (SAQs) usually have a series of acceptable answers that are interpreted as having a single meaning. Even in many objective structured clinical examinations (OSCEs), the scoring rubrics are often associated with either: a) giving points for the initial steps of information gathering, pattern recognition and the final (and typically straightforward) outcome steps of management or,

b) awarding points associated with performing a focused physical examination. For example, in OSCE's associated with standardized patients, if candidates remember to ask the seven cardinal questions of a patient's history and can recall simple characteristics associated with a diagnosis or basic elements of a counseling session, they will often be able to secure enough points to pass the station. Rarely are candidates asked how they are applying that information and more specifically, what key features they are using to develop a working diagnosis, what they are attempting to rule in and out from requested investigations and how they have synthesized a particular diagnosis. Regardless of the format, the scoring keys for these and other forms of assessment are typically derived from "gold standard textbooks or journals" or other evidence based sources such as national medical association statements and discipline specific guidelines. Put simply – it is much easier to create assessments for knowledge-based tests for which there is certainty and very clear evidence. So, why should assessment change?

Current methods of assessment must change because they do not sufficiently reflect the realities of reasoning processes required in clinical medicine. The Canadian Royal College of Physicians and Surgeons (RCPSC) updated CanMeds 2015 framework for physician competencies recognizes this.[35] Concepts of clinical complexity, uncertainty and ambiguity have been acknowledged as important to include within the medical expert role. Similarly, the medical expert role has been updated to reflect some of the complexity of decision-making and clinical reasoning that occurs after the completion of procedures. What is becoming increasingly apparent is that while some real clinical cases are "clear-cut" and basic knowledge, comprehension and application will suffice, there are many cases in which there is a significant degree of *uncertainty*. We propose various situations contribute to this uncertainty including:

1) The amount of currently available information for the case is limited.
2) Key information is missing.
3) Some information is contributory but not “discriminating”.
4) The information available may be weighted differentially.
5) Key features available do not readily “fit” a diagnosis.
6) There is either no evidence or poor evidence for a particular path or course of action.
7) Contextual factors that may be highly relevant to decision-making must enter the equation for successful patient care. These include age, medical acuity and complexity, geographical location of the health care team and the family, available resources (including expertise, time, physical and financial resources), health care system factors, system language and communication issues, cultural beliefs and social circumstances.

Therefore, while knowledge and comprehension continue to be necessary in modern day clinical reasoning (and assessment), they are not sufficient. Clinicians must be able to discern what information is available, what is most relevant and what key pieces are missing. They must *apply* information *in context.* At any given point in time, a clinician must be able to *analyze* the case, *synthesize* a working diagnosis (even if it is not the final diagnosis) and make active decisions regarding investigations and treatment. Although clinicians are always working towards the “goal” of *increased confidence*, they must develop the clinical reasoning processes necessary to manage the initial (and intermittent) situations *uncertainty*. When these “processing” skills receive primary focus, there is a predictable improvement in the desired outcomes of safe, effective and efficient patient care. As stated by Schuwirth and van der Vleuten, medical education research is revealing, “medical

expertise is closely linked to an individual's performance in clinical reasoning."[36,37] Therefore, as medical educators, we have a responsibility to embed these higher order clinical reasoning skills of *application*, *analysis* and *synthesis* into both teaching and assessment and to do this within real world contexts of *uncertainty*.

6.4 Can There Really Be More Than One Right Answer?

In some cases in clinical medicine there is enough information about the patient and sufficient evidence in the literature (or via expert consensus) to adhere to a single, clear path. In other cases, the information available may be "sub-optimal", or, given the information and context available, there may be insufficient evidence to dictate one specific choice. Depending on the case and the context, two (or more) paths may be considered safe and reasonable.

A clinical example is illustrated in Figure 1. The base clinical SCT scenario describes an 11-month-old infant who has a 3-day history of fever, irritability and decreased oral intake. Based on an interpretation of the clinical history and analysis of laboratory investigations, this infant most likely has a urinary tract infection (pyelonephritis). However, at this stage, the evidence is not conclusive. Meningitis is less likely but cannot be fully ruled out. Either way the child requires intravenous (IV) or intramuscular (IM) antibiotics until further information becomes available. Several management options are provided in part one of this case. Based on the reasoning and rationale provided above, giving a single dose of IM Ceftriaxone followed by a return to the emergency department in 24 hours (c) is safe and reasonable. Admitting the child to hospital and initiating IV antibiotics is also safe and reasonable. As the case evolves and more information becomes available (part two), pyelonephritis is confirmed. However, after 48 hours the child still has fever. Therefore, it is inappropriate to either

discharge this patient without further treatment (a) or change to oral antibiotics given the persistent fever (b). Continuing an IV antibiotic (Ampicillin) known to be effective against this infectious organism is necessary. The decision to treat with IV Ampicillin and either wait and observe for 24 hours (c) or conduct further tests (d) are slightly different options but both are acceptable paths. Some children (especially infants) may still have fever on Day 2 of this type of infection, but for the majority of children the fever will have resolved by Day 3. As such, option "c" is reasonable. However, it is also possible that the child has an abscess or (very remotely) has meningitis; therefore, a decision to pursue an abdominal ultrasound and/ or a lumbar puncture (to help rule these conditions out) is also acceptable. This is a classic example where a clinician at each stage has two safe and reasonable options and he/she may select either one (as did the panel of experts for this case). Such situations occur frequently in clinical medicine. Rather than avoiding these clinical situations in the domain of assessment, we have an opportunity to embrace them. We propose that respecting the possibility of the existence of "more than one acceptable path/ correct answer" reflects *clinical reality* and will ultimately make the assessment of clinical reasoning more valid. More specifically, acknowledging candidate responses that are acceptable and progressively rewarding responses that might be considered "most optimal" seems to be more representative assessment of the types of clinical decisions often required by trainees and practicing physicians.

6.5 Current Barriers and Potential Strategies

What are the current barriers to creating assessments that target decision-making in contexts of uncertainty? The first barrier is ***acknowledging situations of uncertainty*** in clinical medicine by all stakeholders involved with assessment. This includes candidates, examiners, test developers, medical

schools, program directors, certifying bodies and the general public. By exposing and highlighting these clinical situations, we come closer to accepting that gray zones will always exist and it is better to acknowledge, confront and manage these situations in formative and summative assessment rather than steering away from them.

The second barrier is that a ***clear framework*** must be developed to identify the progressive levels of clinical reasoning skills needed to manage varying degrees of uncertainty in clinical decision-making. In the early phases, trainees need to be able to identify "key features" embedded within a case. Then they need to apply appropriate "weight" to each of these features within the context of the particular case (as some features will more heavily influence a case at any given point in time). That weighting may be influenced by the quality and the reliability of the information being received [i.e., a sign or symptom or other point (i.e., allergy) on history-taking, a laboratory investigation, a treatment option]. The next step is a critical one. The trainee must evaluate not just the medical patient factors but also the contextual factors that "color the picture" of every clinical case. The textbook case is the "prototype" case but each real case is an "exemplar" with typical and atypical features as well as additional contextual variables that are entrenched and must be considered.[38,39] Teasing out the specific patient case factors with the specific contextual factors requires the clinical reasoning skill of *analysis*: being able to select the most relevant medical and contextual aspects of a case, determining which ones will predominate and synthesizing what this means in terms of clinical decision-making. Using this framework (or another one) allows those creating assessments to build valid cases and questions that help test and measure the progression of the skill of clinical reasoning in contexts of uncertainty. Creation of such a framework is a necessary first step for both test development and assessment.

A third barrier that is very real and extremely relevant to certification bodies relates to the ***defensibility of the assessment method***. In each instance of assessment "an interpretative argument must be built for which evidence is collected in support of proposed inferences."[40] More specifically, "the interpretation of scores must be linked to a network of theory, hypotheses and logic which are presented to support or refute the reasonableness of the desired interpretations."[41] On this point there appears to be some significant debate in the clinical reasoning literature. One criticism of the SCT method is that the scoring method is based on the "cumulative response" of a panel of experts rather than a "consensus response." Charlin et al. contend that, particularly in areas of uncertainty, the cumulative response of the expert panel helps to illuminate acceptable responses including modal responses and responses near to the modal response.[42,43] The SCT method supports two of the core *principles* we feel are paramount to the assessment of modern day clinical reasoning: testing in contexts of uncertainty and respecting the possibility that there may be more than one acceptable path (right answer). A third advantage is that this method allows further discrimination along the continuum of clinical reasoning skill as the examinee does not automatically score a one or a zero for any given question (as in MCQ) but may receive partial credit depending on the frequency of that response by members of the expert panel. In this realm, each individual expert response is considered *valid* and is respected as such. While this belief has tremendous merit, there are at least three risks in taking this view. The first relates to the principle of "case specificity", which states that just because a clinician is an expert and up to date in one area of specialization, does not guarantee that this is consistent across the domain.[44,45] Not all knowledge is maintained uniformly over time and new knowledge is constantly being created. Secondly, other than being "actively practicing" physicians, there are no clear criteria for serving as a panel member. A third risk is that, although well-identified

individual experts may have valid individual responses, some degree of peer review of the scoring rubric would be valuable. To address these three vulnerabilities, the criteria to serve as a panel member could be more robust. In addition to satisfying a minimum number of years of experience, standards could be stipulated regarding the volume (average number of hours per week), intensity (number of weeks per year) and specificity of the physician's clinical practice (main types of patient populations served – this can vary even within a specialty). Secondly, for each given case within an SCT, panel members are able to identify if they feel comfortable in serving as "an expert" for that case including being familiar with current evidence based medicine in the topic area. If they are – they proceed with responding to that case; if not they move on to the next case. This modification may require a slightly larger panel of experts be recruited for the assessment – however, more panel members may be attracted if this option exists. Thirdly, and *especially* in areas of uncertainty it may be more defensible to create SCT cases and questions that still support the standard process of peer review – not necessarily to obtain consensus (as by definition there is no clear consensus or evidence for a single answer) but rather to determine which responses are still safe and reasonable and which options are inappropriate or even dangerous. It is proposed that these modifications may enhance the validity and the reliability of the judgments that are made by the SCT panel of experts and thereby help to make this method of clinical reasoning assessment more defensible, especially in certifying examinations.

A fourth barrier facing meaningful assessment of clinical reasoning in contexts of uncertainty relates to ***feasibility***. In both training assessments and summative evaluations, medical education leaders must be mindful of a multitude of variables: [46-48]

1) Who will develop the assessment? What expertise is required? Where and how will this occur? What is the cost?
2) Where will the assessment take place? Is the location available, easily accessible, secure, and appropriate for the assessment method? Can the location accommodate a large number of candidates? What is the cost of the facility?
3) When will the assessment take place? How long does it take? Does this fit reasonably well With the availability of candidates, examiners and administrators?
4) Does the administration of the assessment require candidates or examiners to incur costs related to travel, accommodation or meals?
5) Does the timing, location, number of candidates and examiners necessitate the creation of more than one examination (i.e., Exam A and Exam B)? Does this threaten/create risk for unintended test exposure/contamination? Is sequestering of candidates necessary?
6) What is the time and financial cost to develop, administer, score and distribute results of the assessment?

Currently, there are few targeted methods for formative or summative assessment of clinical reasoning in graduate training. Some efforts are being taken in some local and national programs to embed some of this assessment in OSCE examinations. While the OSCE can be a very natural and appropriate place to conduct some aspects of clinical reasoning assessment, it is difficult to conduct sufficient sampling due to other competencies also tested by the OSCE method (i.e., data gathering skills, communication skills, manager skills), as well as the limited number of OSCE stations that can be administered in any one OSCE examination. In addition, OSCEs are very expensive and time consuming.[13] In contrast, the SCT method is an exceptionally feasible method for the assessment of

clinical reasoning. Perhaps the most significant requirement is the need for three test developers, ideally all with significant clinical experience in the field as well as training and experience in test development. The cost in terms of time and money is similar or less to the cost associated with MCQ or SAQ assessments. Test administration, tracking and scoring occurs online via a secure website and is conducted simultaneously for an unlimited number of candidates in any location. This means that only one version of the test needs to be developed – a significant advantage. Regarding location and equipment, most medical schools can accommodate multiple students in a computer lab for a 90-minute testing period. There are no other costs to the participants (candidates or expert panel members). Results can be distributed instantaneously and electronically allowing for immediate feedback. This is tremendously useful for both formative and summative assessments. All of these features make SCT a particularly feasible and attractive method.

We have addressed four current barriers associated with the assessment of clinical reasoning within contexts of uncertainty: acknowledging situations of uncertainty, developing clear frameworks for clinical reasoning assessment in these contexts and the defensibility and feasibility of these assessments. We have also proposed specific strategies to overcome these barriers. These strategies may be helpful in enhancing the reliability, validity, feasibility and acceptability of the SCT method - one scheme specifically designed for the assessment of reasoning in contexts of uncertainty. The fifth criterion to consider when developing a robust assessment is the ***impact on future learning and practice***.[6] Medical educators know that assessment drives learning. We support the pursuit of refining the SCT method, as well as the development of other methods for clinical reasoning assessment because we recognize that as these methods are integrated into formative and summative assessments, trainees will be motivated to learn and excel in this medical expert skill. We anticipate

these effects will aid trainees in making better clinical decisions for their patients over the course of their training and with increasing experience, over the course of their career.

6.6 Conclusions

With advancements in health care worldwide, infants, children and adults are surviving illness and living with increasingly complex health care conditions. Concurrently, medical research continues to push the boundaries of our clinical knowledge, and, as we blaze "new trails" we naturally encounter uncharted territory and uncertainty. During this process, however, we also often discover more than one safe and acceptable approach to managing that uncertainty. We have attempted to highlight the need to incorporate two principles: 1) embrace uncertainty and, 2) acknowledge the potential for more than one acceptable path, into formative and summative assessments of clinical reasoning. We recognize that genuine barriers exist on the pathway to creating such assessments and have proposed specific strategies to address these. These include acknowledging situations of uncertainty, creation of clear frameworks that define progressive levels of clinical reasoning skills, providing validity evidence to increase the defensibility of such assessments, considerations of comparative feasibility with other forms of assessment and consideration of strategies to evaluate the impact of these assessment methods on the future learning and practice. We believe concerted efforts directed towards these key areas could help advance the field of clinical reasoning assessment, improve the clinical care decisions made by current and future physicians, and have positive outcomes for patients and families.

6.7 Figures

Figure 8: The Evolving Style SCT Case – Sample Three

Base Scenario: An 11 month-old infant is seen in the emergency department with a 3-day history of fever (up to 39 degrees Celsius) and irritability. He had one episode of non-bilious vomiting yesterday. He has had no diarrhea. The mother has not noticed any problems with his breathing. She has bathed him daily and not noticed a rash. She reports he has been taking ~500 mls of formula daily and 250 mls of water or juice. His intake of solids has been a little less than usual. On physical examination you note the following:

Vital signs:

Heart rate + 134 beats/minute
Respiratory rate = 28 breaths/minute
O2 saturation = 97% in room air
Temperature = 39.6 degrees Celsius
Blood pressure = 90/55 mmHg

He is inactive during your physical examination. He has dry mucous membranes. His capillary refill time is 3 seconds centrally and peripherally. His physical examination is otherwise unremarkable. Blood work and a capillary blood gas are drawn. A peripheral IV is inserted. He receives a bolus of normal saline (20cc/kg) after which time he voids. Repeat vital signs show a heart rate of 115 bpm, respiratory rate of 24/min, blood pressure of 92/56 mmHg and temperature of 39.5 degrees Celsius. He is now a bit more interactive. His investigations reveal the following:

CBCd:

Hgb = 137 g/L
WBC's = 18.6 x 10^9/L
Neutrophils = 10.4 x 10^9/L
Lymphocytes = 4.8 x 10^9/L
Bands = 0.2 x 10^9/L

Chemistry Panel:

Na = 148 mmol/L
K = 5.5 mmol/L
BUN = 8.1 mmol/L
HCO3 = 14 mmol/L
Creatinine = 62 umol/L

Other Labs:

C-reactive protein = 36
CBG: pH = 7.29, CO2 = 42
(Base deficit = -7)

Urinalysis = 10 – 20 WBC's, positive for nitrites, no RBC's, no protein, 2+ ketones.
A blood culture is pending. A lumbar puncture was not performed.

Part One

Considering the treatment plan of …	**This treatment plan becomes …**				
a) Discharge home, parents to return if they have concerns	-2	-1	0	+1	+2
b) Discharge home, prescribe high dose oral Amoxicillin and arrange to call family once culture results are available	-2	-1	0	+1	+2
c) Give one dose of IM Ceftriaxone, discharge home and request family return to the Emergency Department in 24 hours	-2	-1	0	+1	+2
d) Admit, treat with IV Ampicillin and Cefotaxime (pending 48 hour cultures)	-2	-1	0	+1	+2

Part Two

Case Evolution: After 48 hours of treatment with IV Ampicillin and Cefotaxime, his blood culture is negative.
His urine culture is positive for E.Coli sensitive to Amoxicillin, Nitrofurantoin and Septra. His oral intake (fluids and solids) is improving. His parents report his mood is better although his energy is still low.

His vital signs are as follows:

Heart rate = 92 beats/minute
Blood pressure = 96/56 mmHg
Respiratory rate = 18 breaths/minute
Temperature = 38.6 degrees Celsius
O2 saturation = 98% in room air

His physical examination today is unremarkable.

Considering the treatment plan of …	**This treatment plan becomes …**				
a) Discharge home, parents to return if they have concerns	-2	-1	0	+1	+2
b) Discharge home, prescribe high dose oral Amoxicillin x 10 day and follow-up in one week with the family doctor or pediatrician	-2	-1	0	+1	+2
c) Remain in hospital, continue IV Ampicillin and re-assess in in 24 hours	-2	-1	0	+1	+2
d) Remain in hospital, continue IV Ampicillin and pursue further investigations	-2	-1	0	+1	+2

Scale -2 = Contraindicated, -1 = Not indicated, 0 = Neither contraindicated or indicated, +1 = Indicated, +2 = Strongly indicated (essential)

Chapter Seven

Conclusion

It is anticipated that the findings and insights derived from this research can make genuine contributions towards advancing the field of clinical reasoning assessment. In the first overarching study, we demonstrated that by applying specific strategies to SCT design and construction, PSCT scores were able to discriminate clinical reasoning ability between experienced staff and two distinct training levels (senior and junior residents) in pediatric residency. This study also revealed it is possible for a PSCT to achieve the standard of reliability (Cronbach's α = > 0.80) expected of a summative examination. Post-test survey comments expressed by trainees and experienced staff suggest the PSCT covered a range of difficulty, fairly represented the domain of pediatric acute care medicine and accurately depicted "real-life" clinical decision-making. Participants also expressed a positive overall response to the PSCT experience and believed it would be useful to utilize the SCT method of assessment in the future. Each of the findings above support the validity of PSCT scores as an indicator of clinical reasoning competency. We propose that the SCT method may be further improved by making specific modifications to the test design to increase clarity and more fully delineate participant responses. We also recommend that "case load' be considered as a strategy to further discriminate clinical reasoning efficiency.

Our second study revealed, somewhat surprisingly, that the new evolving type case could also be very effective in distinguishing clinical reasoning ability. We postulate that the validity of the evolving case design may be uniquely buoyed by specific features that enhance realism including evolution of a patient's course over time, intermittent (two stage) holistic assessment and respect for

typical contextual points of clinical decision-making. Our qualitative results suggest that both residents and experts find the evolving type case engaging, more similar to real patient cases and perceive the clinical reasoning process required in the evolving type case to be more analogous to the typical decision making processes inherent in true clinical practice.

Finally, we have identified two philosophical principles inherent in the SCT method that have been reinforced through the course of this research. We fully support the concept that for clinical reasoning assessments to most closely reflect clinical reasoning in 'the real world' it is paramount to embrace uncertainty and to also acknowledge the potential for more than one acceptable path. These are bold but necessary steps if assessment of clinical reasoning and decision-making is to align with clinical reality. We recognize that genuine barriers exist on the pathway to creating increased alignment with formal assessment. We have proposed several strategies to help remove these barriers including: 1) acknowledging situations of uncertainty, 2) creation of clear frameworks that define progressive levels of clinical reasoning skills, 3) collecting validity evidence to increase the defensibility of such assessments, 4) consideration of comparative feasibility and, 5) development of schemes to evaluate the impact of these assessment methods on future learning and practice. We believe the engagement of such strategies will help close the gap between current assessment and the realities of clinical reasoning and decision-making in medicine.

This is an exciting time to be conducting research in the assessment of core physician competencies in medicine. This is especially true for the critical competency of clinical reasoning, a skill that has direct implications for clinical decision-making and the quality and efficiency of patient care. As we embark on the development and refinement of clinical reasoning assessments, we have an opportunity to re-visit and re-define underlying principles to ensure they reflect the higher level

cognitive processes and skills essential to perform safely and effectively. We suggest that in order to most optimally move forward future research efforts should focus on: 1) examining and reporting the number and type of clinical situations that create uncertainty, 2) continued refinement of the SCT method incorporating more stringent criteria for test design, construction, standard setting and scoring, 3) increased utilization and refinement of the SCT evolving style case, 4) embedding SCT into annual formative testing and 5) starting to pilot SCT in summative assessments.

In conclusion, it is hoped that the efforts of this research and the suggestions offered for future research endeavors may help to further advance the SCT method and the field of clinical reasoning assessment. As increasingly relevant methods are integrated into formative assessments and summative evaluations, it is anticipated trainees will become motivated to learn and develop this vital clinical competency. These outcomes are predicted to not only improve the confidence of future clinicians in making decisions in contexts of uncertainty, but also the quality of their decisions for patients – the ultimate stakeholder.

References

1. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving, a ten-year retrospective. Evaluation and the Health Profession. 1990; 13: 5 - 36.
2. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. New England Journal of Medicine. 2006; 355(21): 2217-2225.
3. Van der Vleuten CPM. The assessment of professional competence: Development, research and practical implications. Advances in Health Sciences Education. 1996; 1: 41–67.
4. The Royal College of Physicians and Surgeons of Canada: Can MEDS 2005 framework, p. 2; c2005 [cited 2011 Mar10]. http://rcpsc.medical.org/canmeds/bestpractices/framework_e.pdf
5. Accreditation Council for Graduate Medical Education. ACGME Outcome Project: Enhancing residency education through outcomes assessment. 2008 [cited 2011 Mar21]. http://www.acgme.org/Outcome
6. 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: 189-195.
7. Lubarksy, S. Dory, V. Duggan, P., Gagnon, R. Charlin, B. Script concordance testing: From theory to practice. AMEE Guide No. 75. Medical Teacher. 2013; 35: 184-193.
8. Ruiz JG, Tunuguntla R, Charlin B, Ouslander JG, Symes SN, Gagnon R, Phancao F, Roos BA. The script concordance test as a measure of clinical reasoning skills in geriatric urinary incontinence. Journal of the American Geriatric Society. 2010; 58: 2178-2184.
9. Lemay JF, Donnon T, Charlin B. The reliability and validity of a paediatric script concordance test with medical students, paediatric residents and experienced paediatricians. Canadian Medical Education Journal, June 30, 2010. 1(2): e89 – e95.

10. Kow N, Walters MD, Karram MM, Sarsotti CJ, Jelovsek EJ. Assessing intraoperative judgment using script concordance testing through the gynecology continuum of practice. Medical Teacher. 2014; 36: 724-729.
11. Lineberry M, Kreiter CD, Bordage G. Threats to the validity in the use and interpretation of script concordance test scores. Medical Education. 2013; 47:1175-1183
12. See, KC. Keng, LT, Lim, TK. The script concordance test for clinical reasoning: re-examining its utility and potential weakness. Medical Education. 2014; 48: 1069 – 1077.
13. Bland AC, Kreiter CD, Gordon JA. The psychometric properties of five scoring methods applied to the script concordance test. Academic Medicine. 2005. 80; (4), 395-399.
14. Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, Sauve E, van der Vleuten C. Assessment of clinical reasoning in the context of uncertainty : The effect of variability in the reference panel. Medical Education. 2006 ; 40 : 848-854.
15. 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. Radiotherapy and Oncology. 2009; 4:7.
16. Carrière B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in paediatric emergency medicine: Validity evidence for a script concordance test. Annals of Emergency Medicine. 2009; 53: 647-652.
17. Charlin B, Gagnon R, Lubarsky S, Lambert C, et al. Assessment in the context of uncertainty using the script concordance test: More meaning for more scores. Teaching and Learning in Medicine. 2010; 22 (3), 180-86.

18. Goulet F, Jacques A, Gagnon R, Charlin B, Shabah A. Poorly performing physicians. Does the script concordance test detect bad clinical reasoning? Journal of Continuing Education in the Health Professions. 2010; 30 (3): 161-166.
19. Fournier, JP, Demeester, Charlin B. Script concordance tests: Guidelines for construction. BioMedCentral, Medical Informatics and Decision-Making. 2008. 8-18.
20. Objectives of Training in Pediatrics (2008). Royal College of Physicians and Surgeons of Canada. http://www.royalcollege.ca/cs/groups/public/documents/document/y2vk/mdaw/~edisp/tztest3rcpsced000931.pdf
21. Charlin B, Lubarksy S, Kazatani D, Script Concordance Test Web-site. Center for Pedagogical Applications of Science and Health. Faculty of Medicine, University of Montreal, Montreal, Canada. http://www.cpass.umontreal.ca.
22. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology. 2006; 3: 77 – 101.
23. Charlin B, Brailovsky CA, Leduc C, and Blouin D. The diagnostic script questionnaire: A new tool to assess a specific dimension of clinical competence. Advances in Health Sciences Education. 1998; 3, 51-58.
24. Charlin B, Brailovsky CA, Brazeau-Lamontagne L, Samson L, Leduc C. Script questionnaires: Their use for assessment of diagnostic knowledge in radiology. Medical Teacher. 1998; 20, 567-571.
25. Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, van der Vleuten C. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. Medical Teacher. 2002; 24 (5), 522–527.

26. Brailovsky C, Charlin B, Beausoleil S, Cote S, van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. Medical Education. 2001; 35: 430-436.
27. 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. Medical Teacher. 2004; 26 (4): 326–332.
28. Lubarksy S, Chalk C, Kazitani D, Gagnon R, Charlin B. The script concordance test: A new tool assessing clinical judgment in neurology. Canadian Journal of Neurological Science. 2009; 36: 326-331.
29. Park AJ, Barber MD, Bent AE, et al. Assessment of intra-operative judgment during gynecologic surgery using the script concordance test. American Journal of Obstetrics and Gynecology.2010; 203:240. e1-6.
30. Charlin B, Boshuizen HP, Custers EJ, Feltovich PJ. Scripts and clinical reasoning. Medical Education. 2007; 41(12), 1178-84.
31. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. Medical Education. 2012; 46: 552-563.
32. Huitt W. Bloom et al.'s taxonomy of the cognitive domain. Educational Psychology Interactive. 2011. Valdosta, GA: Valdosta State University.
33. Cooke SR, Beran T, Sandhu A, Lemay JF, Amin H. Development of a method to measure clinical reasoning in pediatric residents: the pediatric script concordance test. Publication pending.
34. Cooke SR, Beran T, Lemay JF, Sandhu A, Amin H. Script concordance testing and the evolving case style – is there a new kid on the block? Publication pending.

35. Frank, JR, Snell L, Sherbino J. The Draft CanMeds 2015 Physician Competency Framework – Competency by Design. Royal College of Physicians and Surgeons of Canada Website. March, 2015. http://www.royalcollege.ca/portal/page/portal/rc/canmeds/canmeds2015.
36. Schuwirth LWT, van der Vleuten CPM. General overview of the theories used in assessment: AMEE Guide No. 57. Medical Teacher. 2011; 33(10), 783-97.
37. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Advances in Health Sciences Education. 1996; 1:41 - 67.
38. Bordage G. Prototypes and semantic qualifiers: from past to present. Medical Education. 2007; 41(12), 1117-21.
39. Norman G, Young M, Brooks L. Non-analytical models of clinical reasoning: the role of experience. Medical Education. 2007; 41(12), 1140-5.
40. Kane MT. An argument-based approach to validity. Psychological Bulletin. 1992; 112(3): 527-535.
41. Downing SM. Validity: on the meaningful interpretation of assessment data. Medical Education. 2003; 37: 830 - 837.
42. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: The script concordance approach. Evaluation and the Health Professions. 2004; 27: 304.
43. Elstein AS, Shulman LS, Sprafka SA. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press. 1978.
44. Barrows H, Norman GR, Neufeld VR, Feightner JW. The clinical reasoning process of randomly selected physicians in general practice. Clinical and Investigative Medicine. 1982; 5, 49-56.

45. Brown C, Ross S, Cleland, J, Walsh K. Money makes the (medical assessment) world go round: The cost of components of a summative final year Objective Structured Clinical Examination (OSCE). Medical Teacher. 2015; 1-7.
46. McAleer S. Choosing Assessment Instruments. A Practical Guide for Medical Teachers. 3rd Edition. Ed. JA Dent and RM Harden. 2009; 318-324.
47. Schuwirth LWT, van der Vleuten CPM. How to design a useful test: The principles of assessment. In Understanding Medical Education: Evidence, Theory and Practice. Edited by T. Swanick. Association for the Study of Medical Education. 2010; 195-206.
48. Norcini J, Anderson B, Bollela V. et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Medical Teacher. 2011; 33(3), 206-14.