**Creative Education, 2016, 7, 814-823**
Published Online May 2016 in SciRes. http://www.scirp.org/journal/ce
http://dx.doi.org/10.4236/ce.2016.76084

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

**Suzette Cooke*, Jean-François Lemay, Tanya Beran, Amonpreet Sandhu, Harish Amin**

University of Calgary, Calgary, Canada
Email: *Suzette.Cooke@ahs.ca, JF.Lemay@ahs.ca, tnaberan@ucalgary.ca, Amonpreet.Sandhu@ahs.ca, Harish.Amin@ahs.ca

Received 18 February 2016; accepted 6 May 2016; published 9 May 2016



## Abstract

**Introduction: The Script Concordance Test (SCT) is an assessment method of clinical reasoning skills. SCT is designed to assess a candidate's ability to reason when faced with decisions encountered in the three phases of clinical decision-making: diagnosis, investigation and treatment. Challenges have been raised related to psychometric properties of SCT scores. Data about acceptability of the SCT method are also needed. Objectives: 1) To examine the validity of a Pediatric Script Concordance Test (PSCT) in discriminating clinical reasoning ability between junior postgraduate year (PGY) 1 - 2 and senior PGY 3 - 4 pediatric residents, and pediatricians, 2) To determine if higher reliability could be achieved by applying specific test design strategies to the PSCT and 3) To explore trainees'/physicians' acceptability of the PSCT. Methods: A 24-case/137 question PSCT was administered to 91 residents from four Canadian training centers. Each resident's PSCT was scored based on the aggregate responses of 21 pediatricians (Panel of Experts (POE)). ANOVA was used to compare across the 3 levels of experience. Reliability was calculated using Cronbach's $\alpha$ coefficient. Participants completed a post-test survey about the acceptability of PSCT. Results: Overall, a statistical difference in performance was noted across all levels of experience, $F = 22.84$ ($df = 2$); $p < 0.001$. The POE had higher scores than both senior (mean difference = 9.15; $p < 0.001$) and junior residents (mean difference = 14.90; $p < 0.001$). The seniors performed better than juniors (mean difference = 5.76; $p < 0.002$). Reliability of PSCT scores (Cronbach's $\alpha$) was 0.85. Participants expressed keen interest and engagement in the PSCT. Conclusions: PSCT is a valid, reliable, feasible and acceptable method to assess the core competency of clinical reasoning. We suggest the PSCT 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.**

*Corresponding author.

**How to cite this paper:** Cooke, S., Lemay, J.-F., Beran, T., Sandhu, A., & Amin, H. (2016). Development of a Method to Measure Clinical Reasoning in Pediatric Residents: The Pediatric Script Concordance Test. *Creative Education, 7,* 814-823.
http://dx.doi.org/10.4236/ce.2016.76084

## Keywords

**Clinical Reasoning, Script Concordance Test, Post-Graduate Medical Education, Assessment**

## 1. Introduction

Competent and experienced physicians utilize clinical reasoning to process information necessary to make effective and efficient clinical decisions (Elstein, Shulman, & Sprafka, 1990; Bowen 2006). There is an assumption that trainees gradually build clinical reasoning skills over the course of medical school and residency training (Van der Vleuten, 1996). 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 test knowledge and comprehension, include multiple-choice questions, short-answer questions and objective structured clinical examinations. 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, 2005) and the American Accreditation Council for Graduate Medical Education in the United States (AAC-GME, 2008) 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 (Charlin et al., 2000). Lubarsky et al. have conducted a comprehensive review of the SCT method (Lubarsky et al., 2013).

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 (Ruiz et al., 2010; Lemay, Donnon, & Charlin, 2010; Kow et al., 2014). 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 (Lineberry, Kreiter, & Bordage, 2013; See, Keng, & Lim, 2014). SCT research has also revealed some inconsistency in reliability with scores ranging between 0.40 and 0.90 (Bland, Kreiter, & Gordon, 2005; Charlin et al., 2006; Lambert et al., 2009; Carrière et al., 2009; Charlin et al., 2010; Goulet et al., 2010). 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 (Carrière et al., 2009; Ruiz et al., 2010; Lemay, Donnon, & Charlin, 2010). 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 by recruiting adequate sample SIZES of residents and staff, clearly defining SCT content, selecting an optimal combination of cases and questions, and implementing consistent standards for scoring.. It was proposed that these outcomes could 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 $\alpha$ 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 $\alpha$ 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 resident assessment.

## 2. Methods

### 2.1. PSCT Design

The Pediatric Script Concordance Test (PSCT) was constructed by three RCPSC pediatricians, each of whom possessed 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. (Fournier, Demeester, & Charlin, 2008). A PSCT "test blueprint" was developed using the RCPSC Pediatrics' "Objectives of Training" (RCPSC, 2008). Cases and questions were intentionally created to: 1) ensure a wide array of clinical cases typical of general pediatric in-patient medicine, 2) target the three primary clinical decision-making situations: diagnosis, investigation and treatment, 3) contain varying levels of uncertainty to accurately represent real life clinical decision-making and 4) reflect varying degrees of difficulty to appropriately challenge trainees across a four-year training program. PSCT cases were designed 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**).

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 (Charlin, Lubarksy, & Kazatani, 2015). 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) within the PSCT. It was proposed that this test design could more closely simulate real clinical situations in pediatric in-patient medicine.

### 2.2. Raw Scores

Resident responses to each question were compared with the aggregate responses of the panel of experts as described by Fournier et al. (Fournier, Demeester, & Charlin, 2008). Using this method, residents received a score that reflects the number of panel members that selected the same response. Individual panel member's scores were computed using the aggregate responses of all panel members and with their own set of responses removed (to protect from any potential positive bias). 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.

### 2.3. Score Transformation

Score transformation for the examinees (residents) was performed in a two-step process as outlined by Charlin et al. (Charlin et al., 2010). 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.

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 5 cm below the costal margin. | -2 -1 0 +1 +2 |
| Polyarteritis nodosa | Magnetic resonance venography show 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 1.** A pediatric SCT case.

## 2.4. Participants

RCPSC pediatricians with a minimum of three years of 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 4 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 cases. An email invitation followed each presentation. Recruitment occurred within two months of data collection. Each participant provided written consent prior to test administration.

## 2.5. PSCT Pilot and Optimization

The PSCT was piloted with three residents and two pediatricians 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: 1) maintenance of the Internet connection to the web-based site and, 2) 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 final version of the PSCT consisted of 24 cases and 137 questions.

## 2.6. 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 principal investigator and a research assistant supervised all test administrations. Each testing session began with a 20-minute orientation including: 1) a review of the agenda for the session, 2) a summary of the SCT concept and on-line testing format, 3) a review of SCT 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 designed to invite participant's feedback on the PSCT examination 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.

## 2.7. 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. (Charlin et al., 2010). 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 $\alpha$ coefficient. 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). Qualitative responses were analyzed by two of the investigators using thematic analysis (Braun & Clarke, 2006). The most frequent themes emerging were identified. Representative quotes for each theme were selected and reported.

# 3. Results

## 3.1. Participant Distribution and PSCT Scores

Participant and PSCT scores are presented in **Table 1**.

## 3.2. Time to Completion

All members of the expert panel completed the PSCT in 90 minutes or less. The range was 57 - 90 minutes. A total of 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.

## 3.3. Score Analysis: Inclusion/Exclusion

The final analysis included a total of 12,163 resident responses and 2877 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.

## 3.4. PSCT Score Analysis

One-way ANOVA, effect size and correlations are displayed in **Table 2**. ANOVA demonstrated 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 and the junior residents and the senior residents scored higher than junior residents. 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 $\alpha$ 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 mean scores of any of the study groups.

## 3.5. Post-Test Survey Responses: The PSCT Experience

All participants completed the post-test 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?" Results are displayed in **Figure 2**.

**Table 1.** Results-participant distribution and PSCT scores.

| | N | Mean Score | Range | SD |
|---|---|---|---|---|
| Junior Residents | 51 (33 PGY1, 17 PGY2) | 65.1 | 29.7 - 80.4 | 10.69 |
| Senior Residents | 40 (23 PGY3, 18 PGY4) | 70.9 | 54.4 - 80.4 | 6.73 |
| Panel of Experts | 21 (Mean 8 years experience) | 80.0 | 68.0 - 87.8 | 5.00 |

**Table 2.** Results-one-way ANOVA, effect size and correlation.

| | Mean Difference | Significance (p value) | Effect Size (Cohen's d) | Correlation Coefficient (r) |
|---|---|---|---|---|
| POE vs. SR | 9.1 | <0.001 | 1.54 | 0.61 |
| POE vs. JR | 14.9 | <0.001 | 4.03 | 0.90 |
| SR vs. JR | 5.8 | <0.002 | 1.18 | 0.51 |

POE: Panel of Experts. SR: Senior Residents. JR: Junior Residents. Overall: $F = 22.84$ (df = 2); $p < 0.001$.

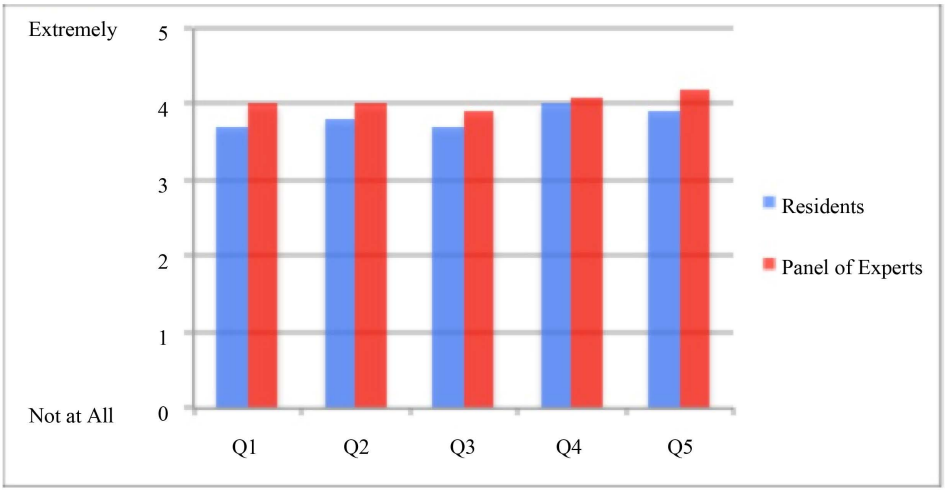


**Figure 2.** The PSCT post-test survey.

Participants also provided 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). Four 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 wish we did this type of thing regularly throughout residency and received feedback to gauge our progress and problem areas." 4) "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."

## 4. 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 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 (Brailovsky et al., 2001; Brazeau-Lamontagne et al., 2004; Carrière et al., 2009; Charlin, Brailovsky, Brazeau-Lamontagne et al., 1998; Charlin, Brailovsky, Leduc, & Blouin, 1998; Lemay, Donnon, & Charlin, 2010; Lubarsky et al., 2009; Park et al., 2010; Ruiz et al., 2010; Sibert et al., 2002). In contrast to the recent SCT "test response validity" concerns of Lineberry et al. (Lineberry, Kreiter, & Bordage, 2013), 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 supported by the results collected in the post-test survey. Participants, including junior and senior level residents as well as less inexperienced and more experienced 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. Participants also uniformly 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 results also revealed that it is possible to achieve a high level of reliability on a PSCT. This outcome may be facilitated when specific strategies in test design and development are applied. In this study, three pediatricians possessing clinical experience in acute-care pediatrics as well as knowledge and experience in test devel-

opment (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). 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 appropriate 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. Firstly, all were required to be certified as specialists in Pediatrics by the RCPSC. Secondly, all POE required a minimum of three years of clinical experience as in-patient staff. It is 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 Canadian 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 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 (Charlin et al., 2007). 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 patient exposure and associated patient volume that comes with increasing clinical experience and have a fewer number of examples to draw on (both typical and atypical cases). Therefore, they 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.

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. Firstly, test developers should be highly selective and precise about the information offered and the options provided. Secondly, 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: 1) use a consistent scale for the entire test regardless of case decision

type (diagnosis, investigation, treatment) and 2) 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 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.

### Limitations

The results of this study should be interpreted in the context of the following characteristics and limitations. First, there were 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 sample 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 POE. A convenience sample of 21 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 POE form the scoring key, there is a risk that how a group of pediatricians practicing 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: 1) the scores of residents from each of the other individual sites as well as with 2) all other sites combined. Since there were no statistical differences, it is suggested that the local POE served as a reasonably unbiased and representative POE 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 4$^{th}$ 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.

## 5. 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, feasible and acceptable 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 and consideration of intentional use of PSCT case load to discriminate clinical reasoning efficiency. We propose that 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.

## Acknowledgements

This research was supported by a grant from the Royal College of Physicians and Surgeons of Canada (Dr. Suzette Cooke-Fellowship Studies in Medical Education). Our research team is highly appreciative of the support and advice provided throughout this project by Dr. Bernard Charlin and the CPASS group (Centre de Pédagogie Appliquée Aux Sciences de la Santé) at the Université of Montréal. The authors wish to thank the pediatric res-

idents, pediatric residency program directors and staff members from the University of Calgary, University of Alberta, University of Saskatchewan and the University of British Columbia. Their willingness to participate in this research has helped to shed light on current and future learning and assessment in residency education. We would also like to express our gratitude to Ms. Linda Beatty for her assistance with preparation of this manuscript.

## Funding

Dr. Suzette Cooke is grateful for the support of the Royal College of Physicians and Surgeons of Canada: 2012 Recipient for a Fellowship for Studies in Medical Education.

## References

American Accreditation Council for Graduate Medical Education (AAC-GME) (2011). ACGME Outcome Project: Enhancing Residency Education through Outcomes Assessment. http://www.acgme.org/Outcome

Bland, A. C., Kreiter, C. D., & Gordon, J. A. (2005). The Psychometric Properties of Five Scoring Methods Applied to the Script Concordance Test. *Academic Medicine, 80,* 395-399. http://dx.doi.org/10.1097/00001888-200504000-00019

Bowen, J. L. (2006). Educational Strategies to Promote Clinical Diagnostic Reasoning. *New England Journal of Medicine, 355,* 2217-2225. http://dx.doi.org/10.1056/NEJMra054782

Brailovsky, C., Charlin, B., Beausoleil, S., Cote, S., & van der Vleuten, C. (2001). 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, 35,* 430-436. http://dx.doi.org/10.1046/j.1365-2923.2001.00911.x

Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology, 3,* 77-101. http://dx.doi.org/10.1191/1478088706qp063oa

Brazeau-Lamontagne, L., Charlin, B., Gagnon, R., Samson, L., & van der Vleuten, C. (2004). Measurement of Perception and Interpretation Skills during Radiology Training: Utility of the Script Concordance Approach. *Medical Teacher, 26,* 326-332. http://dx.doi.org/10.1080/01421590410001679000

Carrière, B., Gagnon, R., Charlin, B., Downing, S., & Bordage, G. (2009). Assessing Clinical Reasoning in Paediatric Emergency Medicine: Validity Evidence for a Script Concordance Test. *Annals of Emergency Medicine, 53,* 647-652. http://dx.doi.org/10.1016/j.annemergmed.2008.07.024

Charlin, B., Brailovsky, C. A., Brazeau-Lamontagne, L., Samson, L., & Leduc, C., (1998). Script Questionnaires: Their Use for Assessment of Diagnostic Knowledge in Radiology. *Medical Teacher, 20,* 567-571. http://dx.doi.org/10.1080/01421599880300

Charlin, B., Brailovsky, C. A., Leduc, C., & Blouin, D. (1998). The Diagnostic Script Questionnaire: A New Tool to Assess a Specific Dimension of Clinical Competence. *Advances in Health Sciences Education, 3,* 51-58. http://dx.doi.org/10.1023/A:1009741430850

Charlin, B., Roy, L., Brailovsky, C., Goulet, F., & van der Vleuten, C. (2000). The Script Concordance Test: A Tool to Assess the Reflective Clinician. *Teaching and Learning in Medicine, 12,* 189-195. http://dx.doi.org/10.1207/S15328015TLM1204_5

Charlin, B., Gagnon, R., Pelletier, J., Coletti, M., Abi-Rizk, G., Nasr, C., Sauve, E., & van der Vleuten, C. (2006). Assessment of Clinical Reasoning in the Context of Uncertainty: The Effect of Variability in the Reference Panel. *Medical Education, 40,* 848-854. http://dx.doi.org/10.1111/j.1365-2929.2006.02541.x

Charlin, B., Boshuizen, H. P., Custers, E. J., & Feltovich, P. J. (2007). Scripts and Clinical Reasoning. *Medical Education, 41,* 1178-1184. http://dx.doi.org/10.1111/j.1365-2923.2007.02924.x

Charlin, B., Gagnon, R., Lubarsky, S., Lambert, C., *et al.* (2010). Assessment in the Context of Uncertainty Using the Script Concordance Test: More Meaning for More Scores. *Teaching and Learning in Medicine, 22,* 180-186. http://dx.doi.org/10.1080/10401334.2010.488197

Charlin, B., Lubarksy, S., & Kazatani, D. (2015). Script Concordance Test Web-Site. Center for Pedagogical Applications of Science and Health. Montreal: Faculty of Medicine, University of Montreal. http://www.cpass.umontreal.ca

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1990). Medical Problem Solving, a Ten-Year Retrospective. *Evaluation and the Health Profession, 13,* 5-36. http://dx.doi.org/10.1177/016327879001300102

Fournier, J. P., Demeester, A., & Charlin, B. (2008). Script Concordance Tests: Guidelines for Construction. *BMC Medical Informatics and Decision-Making, 8,* 18. http://www.biomedcentral.com/1472-6947/8/18 http://dx.doi.org/10.1186/1472-6947-8-18

Goulet, F., Jacques, A., Gagnon, R., Charlin, B., & Shabah, A. (2010). Poorly Performing Physicians. Does the Script Concordance Test Detect Bad Clinical Reasoning? *Journal of Continuing Education in the Health Professions, 30,* 161-166. http://dx.doi.org/10.1002/chp.20076

Kow, N., Walters, M. D., Karram, M. M., Sarsotti, C. J., & Jelovsek, E. J. (2014). Assessing Intraoperative Judgment Using Script Concordance Testing through the Gynecology Continuum of Practice. *Medical Teacher, 36,* 724-729. http://dx.doi.org/10.3109/0142159X.2014.910297

Lambert, C., Gagnon, R., Nguyen, D., & Charlin, B. (2009). The Script Concordance Test in Radiation Oncology: Validation Study of a New Tool to Assess Clinical Reasoning. *Radiotherapy and Oncology, 4,* 7. http://dx.doi.org/10.1186/1748-717x-4-7

Lemay, J.-F, Donnon, T., & Charlin, B. (2010). The Reliability and Validity of a Paediatric Script Concordance Test with Medical Students, Paediatric Residents and Experienced Paediatricians. *Canadian Medical Education Journal, 1,* e89-e95.

Lineberry, M., Kreiter, C. D., & Bordage, G. (2013). Threats to the Validity in the Use and Interpretation of Script Concordance Test Scores. *Medical Education, 47,* 1175-1183. http://dx.doi.org/10.1111/medu.12283

Lubarsky, S., Chalk, C., Kazitani, D., Gagnon, R., & Charlin, B. (2009). The Script Concordance Test: A New Tool Assessing Clinical Judgment in Neurology. *Canadian Journal of Neurological Science, 36,* 326-331. http://dx.doi.org/10.1017/S031716710000706X

Lubarsky, S., Dory, V., Duggan, P., Gagnon, R., & Charlin, B. (2013). Script Concordance Testing: From Theory to Practice. AMEE Guide No. 75. *Medical Teacher, 35,* 184-193. http://dx.doi.org/10.3109/0142159X.2013.760036

Park, A. J., Barber, M. D., Bent, A. E., *et al.* (2010). Assessment of Intra-Operative Judgment during Gynecologic Surgery Using the Script Concordance Test. *American Journal of Obstetrics and Gynecology, 203,* 240.e1-240.e6. http://dx.doi.org/10.1016/j.ajog.2010.04.010

Royal College of Physicians and Surgeons of Canada (RCPSC) (2005). CanMEDS 2005 Framework, p. 2; c2005. http://www.ub.edu/medicina_unitateducaciomedica/documentos/CanMeds.pdf

Royal College of Physicians and Surgeons of Canada (RCPSC) (2008). Objectives of Training in Pediatrics. http://www.royalcollege.ca/cs/groups/public/documents/document/y2vk/mdaw/~edisp/tztest3rcpsced000931.pdf

Ruiz, J. G., Tunuguntla, R., Charlin, B., Ouslander, J. G., Symes, S. N., Gagnon, R., Phancao, F., & Roos, B. A. (2010). The Script Concordance Test as a Measure of Clinical Reasoning Skills in Geriatric Urinary Incontinence. *Journal of the American Geriatric Society, 58,* 2178-2184. http://dx.doi.org/10.1111/j.1532-5415.2010.03136.x

See, K. C., Keng, L. T., & Lim, T. K. (2014). The Script Concordance Test for Clinical Reasoning: Re-Examining Its Utility and Potential Weakness. *Medical Education, 48,* 1069-1077. http://dx.doi.org/10.1111/medu.12514

Sibert, L., Charlin, B., Corcos, J., Gagnon, R., Grise, P., & van der Vleuten, C. (2002). Stability of Clinical Reasoning Assessment Results with the Script Concordance Test across Two Different Linguistic, Cultural and Learning Environments. *Medical Teacher, 24,* 522-527. http://dx.doi.org/10.1080/0142159021000012599

Van der Vleuten, C. P. M. (1996). The Assessment of Professional Competence: Development, Research and Practical Implications. *Advances in Health Sciences Education, 1,* 41-67. http://dx.doi.org/10.1007/BF00596229