Assessing Clinical Reasoning in Pediatric Emergency Medicine: Validity Evidence for a Script Concordance Test

Benoit Carrière, MD, MHPE
Robert Gagnon, MSC
Bernard Charlin, MD, PhD
Steven Downing, PhD
Georges Bordage, MD, PhD

From the Section of Pediatric Emergency Medicine, Department of Pediatrics, Centre Hospitalier Universitaire Sainte-Justine, Montréal, Quebec, Canada (Carrière); the Office of Medical Education, Université de Montréal, Montréal, Quebec, Canada (Gagnon, Charlin); and the Department of Medical Education, University of Illinois at Chicago, Chicago, IL (Downing, Bordage).

Study objective: Clinical reasoning is a crucial skill for all residents to acquire during their training. During most patient encounters in pediatric emergency medicine, physicians and trainees are challenged by diagnostic, investigative, and treatment uncertainties. The Script Concordance Test may provide a means to assess reasoning skills in the context of uncertainty in the practice of pediatric emergency medicine. We gathered validity evidence for the use of a pediatric emergency medicine Script Concordance Test to evaluate residents' reasoning skills.

Methods: A 1-hour test containing 60 questions nested in 38 cases was administered to 53 residents at the end of their pediatric emergency medicine rotation at 1 academic institution. Twelve experienced pediatricians were part of a reference panel to establish the basis for the scoring process.

Results: An optimized version of the test, based on positive item discrimination data, contained 30 cases and 50 questions. Scores ranged from 48% to 82%, with a mean score of 69.9 (SD=11.5). The reliability of the optimized test (Cronbach's α) was 0.77. Performance on the test increased as the level of experience of the residents increased. The residents considered the Script Concordance Test true to real-life clinical problems and had enough time to complete the test.

Conclusion: This pediatric emergency medicine Script Concordance Test was reliable and useful to assess the progression of clinical reasoning during residency training. [Ann Emerg Med. 2009;53:647-652.]

Provide feedback on this article at the journal's Web site, www.annemergmed.com.

INTRODUCTION

Clinical reasoning is a crucial skill for all future physicians to acquire during their training. During most patient encounters in pediatric emergency medicine, physicians and trainees are challenged by diagnostic, investigative, and treatment uncertainties. The clinical supervision of medical students and residents often takes place without direct observation of the trainee's history-taking and physical examination skills. After a brief period of reflection and charting, the trainee reports the findings to an attending physician. From these brief reporting encounters, the attending physician judges the clinical competence of each trainee and reports the judgments on clinical ratings forms. These forms often represent the sole means used to assess clinical reasoning.[1] Although clinical ratings are easy to administer, unobtrusive, and low cost, this assessment system results in many "above-average" ratings, often based on subjective impressions, that do not discriminate well among trainees.[2]

In an effort to improve the validity of the assessment of trainees, both the Accreditation Council for Graduate Medical Education in the United States and the Royal College of Physicians and Surgeons of Canada have asked residency programs to better assess and certify residents' key competencies in becoming qualified physicians, including clinical reasoning. One of the recommendations of the Accreditation Council for Graduate Medical Education advocates the use of more than 1 assessment tool to better evaluate different aspects of competence.

An assessment tool recently developed to assess clinical reasoning, the Script Concordance Test,[3] was implemented in a pediatric emergency medicine program to assess residents at the end of their rotation. The purpose of the present study was to gather validity evidence for the use of Script Concordance Tests in pediatric emergency medicine.

MATERIALS AND METHODS
Study Design

An instrument validation study was conducted. All residents in a pediatric emergency medicine clinical rotation during a 7-month period were asked to complete a Script Concordance Test. The test was administered in a supervised setting during

Table 1. Pediatric emergency medicine Script Concordance Test blueprint.

Topics	Aspects of Clinical Reasoning		
	Diagnosis	Investigation	Treatment
Resuscitation	2	1	2
Respiratory distress	2	1	2
Shock	2	1	2
Arrhythmia	2	1	1
Status epilepticus	2	1	2
Altered level of consciousness	1	2	1
Intoxication	1	1	1
Meningitis	1	2	2
Fever	2	2	2
Gastroenteritis	1	2	1
Abdominal pain	2	2	1
Polytrauma	2	1	2
Fractures, minor trauma	0	2	2
Total questions	20	19	21

the last week of each 4-week pediatric emergency medicine rotation. The Script Concordance Test was completed individually during a maximum 60-minute period after standardized instructions given by a research assistant acting as supervisor. Residents were given the opportunity to ask questions. No financial reward was given to the residents for their participation. They knew that the test would be used only for research purposes and would not be part of their summative evaluation. Approval for the study was obtained from the institutional ethics board, and informed consent was obtained from each participant. On completion of the pediatric emergency medicine Script Concordance Test, every resident completed a brief questionnaire containing demographic data and their ratings (on a visual analog scale) of the degree of difficulty of the test, its clinical relevance, and the time allotted to complete the test. A research assistant coded and compiled all the Script Concordance Test data on a computerized spreadsheet.

The principle behind the Script Concordance Test is to compare examinees' responses (scripts) with those of experienced clinicians using a series of clinical tasks set in specific contexts.[4] The scoring system is designed to measure the concordance between examinee responses (scripts) and those from a panel of expert clinicians.[5] Representative clinical cases from the domain to be tested (excluding unusual or rare cases) are presented to the examinees in short vignettes, each followed

by a series of test questions. For example, to represent the diagnostic challenges of everyday work in pediatric emergency medicine, questions testing the effect of new information on presumed diagnosis in subjects such as meningitis, gastroenteritis, shock, or abdominal pain were used. The indications and usefulness of investigative measures are also tested in other questions because they are a great part of what makes a competent emergency physician. In what situation is a cerebral computed tomography scan needed? When is a lumbar puncture indicated or to be avoided? Finally, the domain of therapeutic decisions is also tested in this Script Concordance Test, with questions testing the indications of treatments, both medical and surgical, in different subjects related to pediatric emergency medicine, such as intoxications, fractures, coma, and arrhythmia.

Two fully qualified Canadian pediatric emergency physicians constructed a test blueprint according to Royal College of Physicians and Surgeons of Canada learning objectives for pediatric emergency medicine. The blueprint contains 13 pediatric emergency medicine topics, listed in Table 1. From the blueprint, a Script Concordance Test was constructed, containing 38 cases, for a total of 60 questions. Sixteen cases contained a single question and 22 cases contained 2 questions each. None of the cases contained more than 2 questions. Three aspects of clinical reasoning were equally represented, that is, 20 questions related to diagnosis, 19 to investigation, and 21 to therapeutic decisions. The test was administered on paper and in French, the working language at the University of Montréal.

Each question contains 3 parts. The first part ("If you were thinking of . . .") contains a diagnosis, investigation, or treatment that is relevant to the case. The second part ("And then you find . . .") contains information, such as a physical sign, a preexisting condition, an imaging study, or a laboratory

Table 2. Example of a Script Concordance Test case used to assess diagnostic reasoning: A 3-year-old girl presents to the ED with important sialorrhea, limited neck movements, and fever for more than 24 hours. Her parents report no recent trauma and no episode of foreign body obstruction.

If You Were Thinking of	And Then You Find	This Hypothesis Becomes
An epiglottitis	An updated vaccination against *Haemophilus influenzae* B	−2−10+1+2
A retropharyngeal abscess	Widened prevertebral soft tissue on a lateral neck radiograph	−2−10+1+2

−2=Ruled out or almost ruled out; −1=less probable; 0=neither less nor more probable; +1=more probable; +2=certain or almost certain.
Reference panel scoring key for first question:
−2: 0 marks; −1: 1.0 mark; 0: 0.2 marks; +1: 0 marks; +2: 0 marks.
Reference panel scoring key for second question:
−2: 0 marks; −1: 0 marks; 0: 1.0 mark; +1: 0 marks; +2: 0 marks.

Table 3. Example of a Script Concordance Test case used to assess investigative reasoning: A 15-year-old girl admits to taking a toxic dose of acetaminophen 30 hours before consultation, after a breakup with her boyfriend. She has no important medical history, and her physical examination shows only a mild tenderness to palpation in the right upper quadrant.

If You Were Thinking of	And Then You Find	This Investigation Becomes
Checking levels of salicylates and other drugs	That the patient denies any coingestion at the time	−2−10+1+2

−2=Totally contraindicated; −1=fairly useless or possibly harmful; 0=neither useful or harmful; +1=useful; +2=indicated or absolutely necessary.
Reference panel scoring key for question:
−2: 0 marks; −1: 0.33 marks; 0: 1.0 mark; +1: 0.16 marks; +2: 0.5 marks.

Table 4. Example of a Script Concordance Test case used to assess therapeutic and resuscitative reasoning: A 2-year-old patient was just intubated after arriving in the ED in cardiac arrest. An adequate endotracheal tube was inserted between the patient's vocal cords. A few minutes later, oxygen saturation is at 85% on 1.0 FiO_2.

If You Were Thinking of Doing	And Then You Find	This Technique Becomes
A needle decompression of a tension pneumothorax	One side of the chest has absent breath sounds on auscultation	−2−10+1+2
A laryngoscopy to check the position of the endotracheal tube	Gargling when auscultating the gastric area	−2−10+1+2

−2=Totally or almost totally contraindicated; −1=not useful, potentially harmful; 0=neither useful or harmful; +1=useful, +2=indicated or absolutely necessary.
Reference panel scoring key for first question:
−2: 0 marks; −1: 0.14 marks; 0: 0.28 marks; +1: 1.0 mark; +2: 0.28 marks.
Reference panel scoring key for second question:
−2: 0 marks; −1: 0.33 marks; 0: 0.16 marks; +1: 0.5 marks; +2: 1.0 mark.

test result, that could have an effect on the first part. The third part ("This hypothesis becomes . . .") uses a 5-point Likert scale (from –2 to +2) for the examinee to indicate what effect the given information (part 2) has on the proposed diagnosis, investigation, or treatment (part 1). Examples of cases related to each domain tested are provided in Tables 2 through 4. The cases are constructed using a "key features" approach,[6] that is, the questions focus on the elements that are most critical or more likely to lead to error in solving the clinical problem.[7] Each question is built to promote reflection before answering.[8] The complete Script Concordance Test in English is available to program directors and medical education researchers on request for secure uses.

Study Setting and Selection of Participants

Residents were recruited from different training levels in various residency programs (ie, family medicine, emergency medicine, radiology, dermatology, and pediatrics) while they were completing a mandatory pediatric emergency medicine rotation at one institution, the emergency department (ED; mean annual visit census of 65,000 patients) at the Centre Hospitalier Universitaire Sainte-Justine in Montréal, affiliated to the University of Montréal. Four to 8 residents were present during each rotation, with every resident working 14 to 16 eight-hour shifts.

To create scoring keys for each question, the Script Concordance Test was submitted to a reference panel of 12 experienced pediatric emergency physicians; 10 to 15 panel members are needed to obtain reliable scores.[9] For each answer to a question, a credit is awarded according to the number of panel members who chose that answer, divided by the modal value for the question. If, for example, 9 panel members (of 12) chose the "+1" option, 2 chose the "+2" option, and 1 chose the "0" option, the credit awarded for an examinee selecting the +1 option would be 1 (9/9), 0.22 (2/9) for the +2 option, and 0.11 (1/9) for the 0 option. For the response options not selected by members of the reference panel, the "–1" and "–2" options in this case, 0 credit is awarded to an examinee selecting those responses. With this scoring procedure, all the questions have the same maximum (1) and minimum (0) values. Case scores were created by adding the question scores and dividing by the number of questions; thus, each case is weighted equally throughout the test. Examinee scores obtained for each case are added to create a total score for the test. In the Script Concordance Test format, a score of 100 would mean that the person has answered each question exactly like the majority of panel members.

The reliability of Script Concordance Test scores in previous studies has always been calculated with the questions as the unit of measurement. Some of these studies used from 5 to 7 questions per case.[10,11] Because of the possible interdependence

Table 5. Mean pediatric emergency medicine Script Concordance Test scores according to levels of residency training.

Level	N	Mean Score	SD	95% Confidence Interval for Mean	Range
PGY-1	21	66.5	8.2	62.8–70.25	48–79
PGY-2	21	68.9	9.9	64.4–73.5	49–84
Seniors	11	75.9	3.9	73.3–78.6	70–82

of questions within cases, the item independence assumption could be violated and would tend to inflate reliability levels.[12] For this study, both methods of calculating reliability will be used and results will be compared. From a theoretical perspective, cases should be used as the basis for calculation to avoid inflating the reliability of the scores.

To optimize the selection of the best items to be used to calculate total test scores, items with negative discrimination, that is, with negative item-total correlations, were discarded. Thus, only items with positive correlations were used to calculate total test scores.

Statistical Analyses

The reliability of the Script Concordance Test scores was calculated with Cronbach's α coefficients. The Spearman-Brown prophecy formula was used to calculate the number of cases that would be needed to obtain a Cronbach's α of 0.80. Statistical analyses were done with SPSS for Windows (version 11.5; SPSS Inc, Chicago, IL). Results were considered statistically significant at the 0.05 level. Sample size was calculated before the study began, with the hypothesis that a moderate correlation of 0.35 would be found between Script Concordance Test scores and global ratings. With PASS software (NCSS Inc, Kaysville, UT), with a power of 0.80 and an α error of 0.05, it was determined that 49 subjects were needed to detect a correlation of 0.35.

Two subject characteristics were analyzed to assess the differences in Script Concordance Test performance among subgroups of residents: (1) training levels (post graduate year-1, post graduate year-2 and "senior" [post graduate year-3 and post graduate year-4 and clinical fellows in pediatric emergency medicine]) residents, and (2) residency programs (ie, family medicine, pediatrics, emergency medicine, and "others"). Descriptive statistics and ANOVA analyses (with Scheffe post hoc analyses) were done. Descriptive statistics were calculated for the 3 visual analog scales (ie, time allotted to complete the task, similarity of the cases with real-life clinical problems, and perceived level of difficulty of the Script Concordance Test).

RESULTS

During the 7-month study period, 53 of 55 eligible residents consented to participate (96%): 38 women (72%) and 15 men (28%). Thirty-four residents were from family medicine (64%), 10 from pediatrics (19%), 3 from emergency medicine (6%), and 6 from other programs (11%; 5 from radiology and 1 from dermatology). Twenty-one residents (40%) were in their first

postgraduate year of training, 21 were PGY-2s (40%), and the 11 others (20%) were in their senior years.

The optimization process of removing negatively discriminating items resulted in the removal of 8 cases containing 10 questions, thus creating a Script Concordance Test composed of 30 cases and 50 questions. The proportional representativeness of the 3 aspects of clinical reasoning remained intact with the optimized test (diagnosis 35%, investigation 31%, treatment 33%). All further reported results are done with the optimized instrument.

The reliability of the scores (Cronbach's α coefficient) was 0.79 with questions as the unit of measurement and 0.77 with cases. With the Spearman-Brown prophecy formula, 36 cases are needed to obtain α reliability coefficient of 0.80.

The correlation between Script Concordance Test scores and global ratings given at the end of each clinical shift by the supervising physician was 0.37 ($P < .01$).

Overall, there was a significant difference in performance across levels of training (F=4.67; df=2, 50; P=.014). Senior residents were stronger than PGY-1s, with a mean difference of 9.4 (Scheffe post hoc test, P=.015). The difference between PGY-2 and senior residents (6.9) was not statistically significant (P=.09); see Table 5. There was no overall significant difference in Script Concordance Test scores when comparing residency programs (F=1.15; df=3; P=.34; power=0.29).

The mean perceived level of difficulty of the Script Concordance Test by the residents was 5.6 on a 10-point scale (10 being most difficult; SD=1.3; median=5.6; range=2.6 to 7.8). The residents overall believed that they had enough time to complete the test (mean=4.7/10 [the lower the better]; range=2.3/10 to 8.4/10); only 4 residents (8.2%) thought that they needed more time. Two thirds of the residents considered that they had enough time, with the majority of the others stating that they had too much time. The mean rating of the representativeness of the Script Concordance Test content to actual situations was 7.4 (10 being best; SD=1.7; median=7.9; range=2.4 to 9.7).

LIMITATIONS

This study used a nonrandom group of residents, a convenience sample.[13] Although this method is less ideal than a stratified random sample, the study setting did not allow for random sampling. However, the group of residents in this study can be considered representative of residents in academic pediatric emergency medicine services in Canada and the United States. For example, the proportion of emergency medicine residents in this study (16%) is similar to that of

residents in a study reported by Chen et al[14] (20%). Family medicine residents in other training programs in Canada also rotate in pediatric emergency medicine for at least 1 month, making it the largest group of residents trained.

As for other threats to internal validity,[13] location was fully controlled by the repeated use of the same testing room, time of day, and testing period during the rotation for all the subjects in the study. Because this study did not affect the final summative evaluation of the residents, it is unlikely that contamination occurred whereby residents would divulge the content of the test to other residents.

The small sample size likely affected the findings when subgroup analyses were done. The composition of the subjects group, with residents from different training programs being recruited into the study, might have introduced bias. If residents from only family medicine had been recruited, albeit from different training levels, the homogeneity of the sampled subjects would have increased.

DISCUSSION

The results from this study contribute positively to a growing body of literature on the Script Concordance Test approach. Previous studies addressed issues of validity, reliability, feasibility, and applicability in different clinical disciplines and contexts.[3,5,8,15] The development of a Script Concordance Test for pediatric emergency medicine and the positive validity evidence gathered in the present study provide additional support for the use of the Script Concordance Test.

For any "written assessment, documentation of validity evidence related to the content tested is the most essential."[16] The optimized test blueprint used in the present study contained 12 topics, equally weighted for 3 aspects of clinical reasoning, namely, diagnosis, investigation, and treatment. For example, "resuscitation" was assessed in 5 questions (10% of the weight of the whole test), and "fever" was assessed in 4 questions (8% of the test). Similarly, the 3 aspects of clinical reasoning were almost equally tested: 17 items related for diagnosis, 15 for investigation, and 16 for therapeutic decisions.

According to Downing,[16] "internal structure, as a source of validity evidence, relates to the statistical or psychometric characteristics of the examination questions or performance prompts, the scale properties—such as reproducibility and generalizability, and the psychometric model used to score and scale the assessment." Although slightly increased, the use of question-based scores did inflate the reliability coefficient, from 0.77 to 0.79. The almost identical question-based and case-based reliability coefficients can be explained by the fact that many of the Script Concordance Test cases (12/30) were single-question cases and that the other cases (18/30) had 2 questions. In other words, the scores for more than a third of the test (12 single-question cases) were identical when the reliability coefficients were calculated. In any event, cases should be the basis for calculating reliability to satisfy the assumption of item independence. A recent study by Norman et al[17] also showed that reliability for case-derived scores is optimal with 2 to 3 questions per case. Fewer questions result in lower

reliability levels, whereas more questions do not improve reliability and do not provide useful psychometric information. Overall, internal structure validity evidence was well documented in this study, representing acceptable reliability coefficients (0.77, 0.79) for a locally developed test (compared to high-stakes national examinations).[12] With 36 well-constructed cases, the Script Concordance Test scores would reach acceptable levels of reliability, which is crucial if a test such as the pediatric emergency medicine Script Concordance Test is to be used as part of a summative evaluation. The construction of a 30- to 36-case pediatric emergency medicine Script Concordance Test is feasible for any given residency program.

The difference in performance between training levels speaks in favor of the validity of the pediatric emergency medicine Script Concordance Test. These are the first results showing that Script Concordance Test scores increase with levels of training among residents. The fact that more experienced residents perform better than less experienced ones is logical and desired. Previous studies using Script Concordance Tests have shown similar results but with medical students from different levels (students versus clerks versus residents).[10,11,18,19] The type of residency program did not influence performance on the pediatric emergency medicine Script Concordance Test.

The range of scores on the Script Concordance Test was fairly wide, with low scores approximately 45% and high scores approximately 85%. Thus, the use of Script Concordance Test scores to assess clinical reasoning could contribute positively to the discrimination power of clinical reasoning skills among residents compared to clumped results on other instruments (eg, all above average on global rating scales).

The residents' opinions concerning the pediatric emergency medicine Script Concordance Test contributed positively to the consequential validity of the test. The time allotted to complete the task was adequate and the level of difficulty was in the medium range. The majority of residents considered that the pediatric emergency medicine Script Concordance Test contained problems that were representative of actual clinical situations.

In conclusion, the scores from this pediatric emergency medicine Script Concordance Test were reliable, especially for a locally developed assessment tool, and test construction was feasible for a residency program. Increasing Script Concordance Test performance followed training levels. A revised pediatric emergency medicine Script Concordance Test is being developed and will be offered as a formative test to students and residents in the study institution. Future studies of the Script Concordance Test approach need to focus on setting of pass-fail if Script Concordance Test scores are to be used for summative evaluations.

Address for correspondence: Benoit Carrière, MD, MHPE, Section of Pediatric Emergency Medicine, Department of Pediatrics, CHU Ste-Justine, 3175 Côte Ste-Catherine, Montreal, QC, Canada; 514-345-4931 ext 2294, fax 514-345-2358; E-mail benoitcarriere@mac.com.

REFERENCES

1. Bandiera GW, Morrison LJ, Regehr G. Predictive validity of the Global Assessment Form used in a final-year undergraduate rotation in emergency medicine. *Acad Emerg Med.* 2002;9:889-895.
2. Gray JD. Global rating scales in residency education. *Acad Med.* 1996;71(1 suppl):S55-63.
3. Charlin B. *Standardized Assessment of Ill-defined Clinical Problems: the Script Concordance Test* [PhD thesis]. Maastricht, The Netherlands: University of Maastricht; 2002.
4. Charlin B, Boshuizen HP, Custers EJ, et al. Scripts and clinical reasoning. *Med Educ.* 2007;41:1178-1184.
5. Charlin B, Van Der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. *Eval Health Prof.* 2004;27:304-319.
6. Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. *Acad Med.* 1995;70:194-201.
7. Caire F, Sol J-C, Moreau J-J, et al. Auto-évaluation des internes en neurochirurgie par tests de concordance de script (TCS): processus d'élaboration des tests. *Neurochirurgie.* 2004;50:66-72.
8. Charlin B, Roy L, Brailovsky C, et al. The Script Concordance Test: a tool to assess the reflective clinician. *Teach Learn Med.* 2000; 12:189-195.
9. Gagnon R, Charlin B, Coletti M, et al. Assessment in context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Med Educ.* 2005;39:284-291.
10. Charlin B, Brailovsky C, Leduc C, et al. The Diagnosis Script Questionnaire: a new tool to assess a specific dimension of clinical competence. *Adv Health Sci Educ.* 1998;3:51-58.
11. Sibert L, Charlin B, Corcos J, et al. Stability of clinical reasoning assessment results with the Script Concordance Test across two different linguistic, cultural and learning environments. *Med Teach.* 2002;24:522-527.
12. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ.* 2004;38:1006-1012.
13. Fraenkel JR, Wallen NE. *How to Design and Evaluate Research in Education.* 4th ed. San Francisco, CA: McGraw-Hill; 2000.
14. Chen EH, Shofer FS, Baren JM. Emergency medicine resident rotation in pediatric emergency medicine: what kind of experience are we providing? *Acad Emerg Med.* 2004;11:771-773.
15. Charlin B, Desaulniers M, Gagnon R, et al. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teach Learn Med.* 2002; 14:150-156.
16. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ.* 2003;37:830-837.
17. Norman G, Bordage G, Page G, et al. How specific is case specificity? *Med Educ.* 2006;40:618-623.
18. Brailovsky C, Charlin B, Beausoleil S, et al. 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. *Med Educ.* 2001;35:430-436.
19. Brazeau-Lamontagne L, Charlin B, Gagnon R, et al. Measurement of perception and interpretation skills during radiology training: utility of the script concordance approach. *Med Teach.* 2004;26: 326-332.