Experts' responses in script concordance tests: a response process validity investigation

Matthew Lineberry,[1] Eduardo Hornos,[2] Eduardo Pleguezuelos,[2] Jose Mella,[2] Carlos Brailovsky[3] & Georges Bordage[4]

CONTEXT The script concordance test (SCT), designed to measure clinical reasoning in complex cases, has recently been the subject of several critical research studies. Amongst other issues, response process validity evidence remains lacking. We explored the response processes of experts on an SCT scoring panel to better understand their seemingly divergent beliefs about how new clinical data alter the suitability of proposed actions within simulated patient cases.

METHODS A total of 10 Argentine gastroenterologists who served as the expert panel on an existing SCT re-answered 15 cases 9 months after their original panel participation. They then answered questions probing their reasoning and reactions to other experts' perspectives.

RESULTS The experts sometimes noted they would not ordinarily consider the actions proposed for the cases at all (30/150 instances [20%]) or would collect additional data first (54/150 instances [36%]). Even when groups of experts agreed about how new clinical data in a case affected the suitability of a proposed action, there was often disagreement (118/133 instances [89%]) about the suitability of the proposed action before the new clinical data had been introduced. Experts reported confidence in their responses, but showed limited consistency with the responses they had given 9 months earlier (linear weighted kappa = 0.33). Qualitative analyses showed nuanced and complex reasons behind experts' responses, revealing, for example, that experts often considered the unique affordances and constraints of their varying local practice environments when responding. Experts generally found other experts' alternative responses moderately compelling (mean ± standard deviation 2.93 ± 0.80 on a 5-point scale, where 3 = moderately compelling). Experts switched their own preferred responses after seeing others' reasoning in 30 of 150 (20%) instances.

CONCLUSIONS Expert response processes were not consistent with the classical interpretation and use of SCT scores. However, several fruitful and justifiable alternatives for the use of SCT-like methods are proposed, such as to guide assessments for learning.

Medical Education 2019
doi: 10.1111/medu.13814

[1]Zamierowski Institute for Experiential Learning, University of Kansas Medical Center and University of Kansas Health System, Kansas City, Kansas, USA
[2]Practicum Institute of Applied Research in Health Sciences Education, Madrid, Spain
[3]College of Family Physicians of Canada, Toronto, Ontario, Canada
[4]Department of Medical Education, College of Medicine, University of Illinois at Chicago, Chicago, Illinois, USA

Correspondence: Matthew Lineberry, Zamierowski Institute for Experiential Learning, University of Kansas Medical Center and University of Kansas Health System, Sudler Hall G005, 3901 Rainbow Boulevard, Kansas City, Kansas 66160, USA.
Tel: 00 1 913 588 0422; E-mail: mlineberry@kumc.edu

INTRODUCTION

Educators and administrators seek to assess clinical reasoning beyond the memorisation of facts, towards the application of more advanced knowledge and the skills needed in complex cases.[1] The script concordance test (SCT) was meant to accomplish this by assessing the interpretation of clinical data in cases without clearly known correct answers.[2] As a by-product of the method's approach to establishing a scoring key, SCT research has quantified how much expert clinicians disagree about clinical data interpretation in such cases. However, rather than assuming all such variations are clinically valid, as is implied by the method's classical scoring technique ('aggregate scoring'), we sought to explore these variations and to understand the *response processes* underlying them.

Script concordance test cases begin with a patient *case description*, often in written form and perhaps with multimedia details. Next, several *proposed actions* are listed, including diagnoses, investigations and treatments that might be undertaken for that case. Then, new clinical *data* are presented, after which the examinee must report his or her *post-data belief change* in the suitability of each proposed action. Often this is measured on a 5-point scale, indicating belief that the new data make the proposed action 'much less suitable' (−2) to 'much more suitable' (+2), with the midpoint (0) indicating that the data have no bearing.

Script concordance test cases are meant to elicit experts' case-relevant *illness scripts*,[2] which are instantly and effortlessly activated complex knowledge structures in memory. In experts, these knowledge structures have been assembled over time, often without conscious thought or effort, from repeated exposure to associations between signs, symptoms, disease courses and other contextual details seen during real-world experiences with patients.[3] The methodology of the SCT calls for the writing of cases with relatively few details, purportedly to best activate scripts in memory.[4] As SCT developers have assumed there are no empirically known correct answers for SCT items, rather than using clinical research or expert consensus conferences to set the scoring key, the test is instead given to a panel of experts who complete the test independently.[5] Each possible answer is considered correct according to how many experts endorse it (termed 'aggregate scoring'). The most frequently chosen response is deemed 100% correct; responses chosen less frequently are awarded partial credit.

Recently, several studies have suggested that the use of experts' responses for scoring in this way is problematic. Lineberry et al.[6] demonstrated three major validity flaws: (i) aggregate scoring often causes scoring keys to be illogical; (ii) SCT research reports markedly overestimate score reliability as a result of a lack of consideration for most ways in which unreliability can manifest in the scoring system, and (iii) a credit anomaly occurs whereby more extreme answers (e.g. 'much more suitable') earn less credit on average as a result of limits on how partial credit can be distributed. Kreiter[7] further noted that as SCTs do not consider what respondents believed the suitability of a proposed action to be *before* they received new data (their 'pre-data belief'), it is impossible to clearly interpret their post-data belief change. Figure 1 illustrates two examples of how this can lead to misinterpretations, whereby experts' agreement or disagreement in their post-data belief change is the opposite of whether they agree about the proposed action's final suitability. Additionally, Ahmadi et al.[8] compared experts' SCT responses against clinical evidence, found that experts' responses were discordant with clinical evidence for 73% of items and questioned whether clinical evidence is truly unavailable for other SCTs and whether all experts' responses reflect a valid divergence of opinion.

Beyond these issues, several scholars have called for investigation into whether experts' and examinees' *response processes* are consistent with the test's validity argument.[9–12] For instance, Power et al.[13] prompted examinees to justify their answers on several SCT cases and found that examinees often: (i) held incorrect rationales for 'correct' SCT answers; (ii) held correct rationales for 'incorrect' answers, and (iii) interpreted questions in unintended ways. To complement and extend this work, we sought to focus on the expert scoring panel, rather than the examinees, in a *response process validity investigation*.[14] Our research question was: What are the reasons for expert panellists' apparent agreement or disagreement on SCT items?

Borsboom et al.[15] challenged validity researchers to specify a *theory of response behaviour* that delineates which response processes are consistent versus inconsistent with one's validity argument. Currently, response theory lags far behind assessment practice[16] and is rarely investigated in health professions education broadly[17] or in SCT scholarship.[9] Thus, we tested a posited theory of response process grounded in the classical SCT interpretation and use argument, which is that:

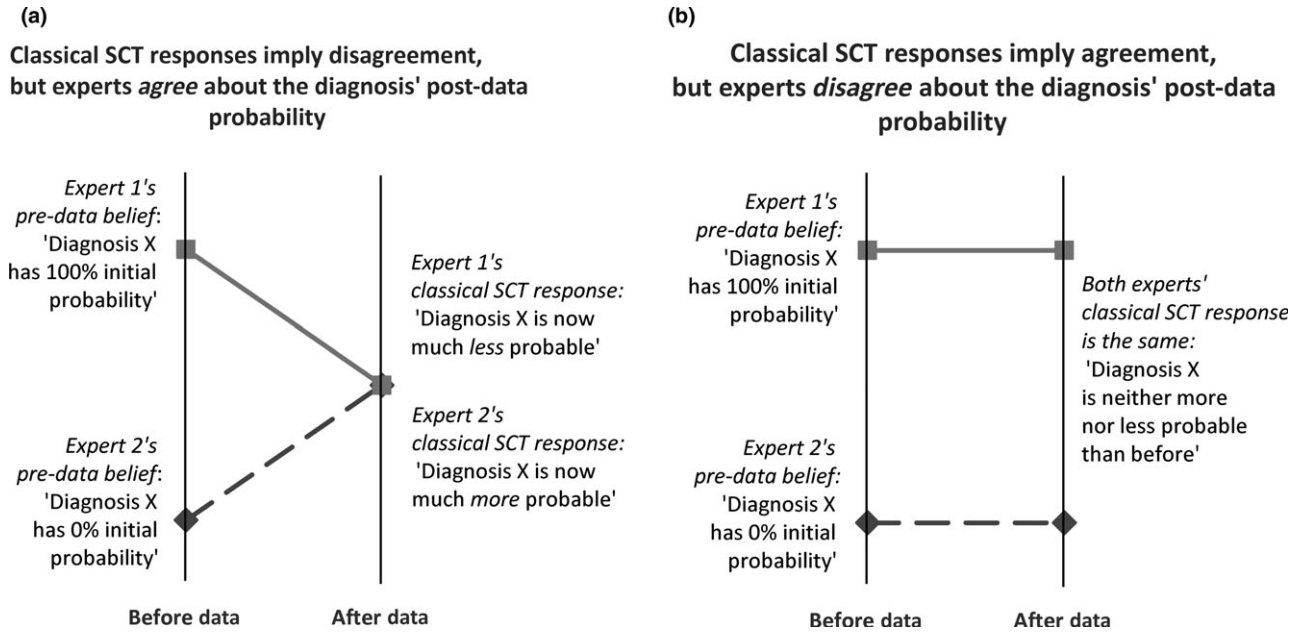

Figure 1 Two hypothetical instances with potential for misinterpretation of script concordance test (SCT) responses if experts' pre-data beliefs are not known and considered: SCT responses may mask underlying agreement (Fig. 1a) or may fail to reflect underlying disagreement (Fig. 1b)

1 the SCT has *psychological fidelity* to actual clinical reasoning and thus activates experts' well-learned illness scripts;
2 expert respondents' post-data belief changes can be interpreted even without knowing their *pre-data beliefs*;
3 expert respondents are neither guessing nor answering items inconsistently;
4 different expert responses reflect genuine clinical disagreement, and
5 experts hold their stated beliefs firmly.

METHODS

The study design is a prospective, mixed-methods survey of experts' beliefs for a set of SCT cases on which experts previously showed disagreement.

Participants

A total of 10 experts previously completed an early version of the Practicum Script in gastroenterology, a continuing professional development programme designed to allow physicians from several specialties to practise clinical reasoning through weekly patient cases, which were originally based on the classical SCT methodology.[18] All agreed to participate. All were board-certified gastroenterologists in Argentina, with at least 10 consecutive years of gastroenterology clinical experience, involved in teaching roles in residency programmes and in postgraduate continuing medical education in the previous 5 years, and spent at least 50% of their working time in patient care both currently and in the previous 5 years. In total six experts were in private practice and four experts were in public practice, all were at teaching hospitals. All were native Spanish and fluent English speakers. There were no participation incentives. The Institutional Review Board at the University of Illinois at Chicago approved the protocol.

Measures and procedures

Cases were developed by a separate group of 10 specialists in gastroenterology with the same professional criteria as the participating experts. Case development was guided by a blueprint, previously defined according to the needs of the specialty. Case writers were asked to provide complex and controversial cases drawn from the histories of real patients according to SCT development guidelines[5] and were given full-day workshops for training in the methodology. Cases were reviewed by at least two clinical specialists before being completed by the expert panel.

Experts completed an online questionnaire developed in English, translated into Spanish by two bilingual members of the research team (EP and

JM) and verified by a translation professional. A total of 15 SCT items were selected and stratified by the type of proposed action (diagnosis, investigation, treatment) so that each type of proposed action was associated with five items. Each item was based on a unique case. Experts first reported their pre-data beliefs about the proposed action, along with their level of confidence in and rationale for the belief (the latter in free text). Secondly, the new data were presented and experts reported any changes in post-data belief, confidence and rationale. If the change in post-data belief was different from that reported 9 months earlier, the expert was shown the earlier answer and asked what he or she believed had caused the answer to change. Finally, experts viewed summaries of other experts' answers and rationales for post-data belief changes, along with clinical evidence associated with each position; at least one article consistent with each position was available for all cases. Then, the experts rated how compelling each rationale was, indicated whether they would like to change their post-data belief change responses, and shared general comments about the case or SCT process. Experts completed 'Diagnoses' cases first, 'Investigations' a week later, and 'Treatments' a week later. An example case is available in Appendix S1.

Analyses

Quantitative data are reported descriptively and analysed according to data distributions. Qualitative data were open-coded initially by two authors (ML and GB), independently and then in concert, until an initial coding scheme was derived. All other authors (EP, CB, JM and EH) reviewed the coding scheme and initially assigned codes, and then proposed edits and additions. After two rounds of revision and extensive discussion, we reached a consensus coding scheme and set of codes. Reflecting on how our personal orientations might influence our interpretations, our team featured a mix of clinical and educational expertise, with a range of postures towards the SCT but with all investigators suspecting the method may have unintended issues. Our posited theory of response process was established prior to reviewing any data.

RESULTS

There were no missing data for selected-response items. For the three open-ended items, experts gave extensive rationales, with median word counts of 29, 42 and 39 English words, respectively.

Pre-data beliefs

Experts held varied pre-data beliefs. Although in 66 of 150 (44%) instances experts indicated that they would ordinarily consider the proposed action at that point in their reasoning, in 30 of 150 (20%) instances, experts indicated 'No, I would not order or consider *the proposed action*' at all at that point in their clinical reasoning, and in 54 of 150 (36%) instances, they indicated they would first collect additional clinical data before evaluating actions.

All experts indicated at least 'moderate' confidence in their pre-data beliefs ('3' on a 5-point scale; 40/150 [27%] 'moderately confident'; 96/150 [64%] 'very confident', 14/150 [9%] 'completely confident'). Codes for pre-data belief rationales are summarised in Table 1. Many experts specified the additional clinical data they would seek (51 instances) and additional proposed actions they had in mind (27 instances). For example, Expert 10 on Case 3330-1 stated:

> Ultrasound is not a very effective method to define 'antral deformation' [referring to the case description's details]. In my opinion, the patient needs an endoscopic and/or radiological examination; depending on the results, GIST [gastrointestinal stromal tumour, the proposed diagnosis] might be among the possible differential diagnoses, along with gastric cancer and diffuse gastric disease.

In six instances, experts indicated a desire to clarify the patient's story. For example, Expert 5, on Case 3351-1, wrote:

> . . . some of the . . . pain interpreted as renal colic might have been undiagnosed acute pancreatitis, which is plausible if nobody thought of it at the time.

In three cases, experts noted that proposed actions may be suitable *later* but not now (six instances); Expert 7, Case 3350-1, stated:

> She has severe acute pancreatitis . . . this [proposed action] would be a study to do after she has gotten over her acute problem.

Post-data belief changes

Confidence in post-data belief changes was similar to that for pre-data beliefs (49/150 [33%] 'moderately confident'; 91/150 [61%] 'very confident'; 10/150

Table 1 Codebook: rationales for pre-data beliefs

Code	Code description	Instances	Experts making at least one such statement (out of 10 experts), *n*	Example statement
1	Participant basically agrees that the case is consistent with the proposed action at this point in the case	67	10	'Doing pH monitoring would tell us whether the reflux is acidic or alkaline.' (Expert 3, Case 3296-3)
2	Participant wants more information (e.g. laboratory tests, history)	51	10	'Before pH monitoring, I would evaluate the oesophagus with a new endoscopy and see what happened with the treatment and whether he has Barret's oesophagus (like his father) under the oesophagitis.' (Expert 8, Case 3296-3) (also coded as Code 3)
3	Participant believes the proposed action has relatively low probability or utility at this point in the case; would consider other actions first	49	10	'Since he has oesophageal erosions, the diagnosis of GORD [gastro-oesophageal reflux disease] already confirms the diagnosis. I consider pH monitoring unnecessary.' (Expert 7, Case 3296-3)
4	Participant has other actions in mind, without implying higher or lower probability or utility	27	8	'As it is the intrahepatic biliary duct that is dilated, it would be good to do ERCP [endoscopic retrograde cholangiopancreatography].' (Expert 8, Case 3353-1)
5	Participant notes some case details are at least partially inconsistent with the proposed action	12	9	'This diagnosis is common 24 hours after ERCP, fever and pain support it, although melena doesn't fit.' (Expert 8, Case 3352-5)
6	Participant wants to clarify or validate the patient's story or case details	6	5	'I would like to know where the colonoscopy was done and its quality.' (Expert 1, Case 3299-2) (also coded as Code 10)
7	Participant believes the proposed action should not be pursued until the patient is stabilised	6	5	'The chest X-ray shows pleural effusion. The severity of the pancreatitis should be defined and it should be appropriately managed before doing a study to determine its cause.' (Expert 2, Case 3350-1)
8	Participant believes there are multiple partially correct answers	5	5	'Answers B [*less likely*] and C [*neither more nor less likely*] are correct. In patients with angiodysplastic bleeding that cannot be controlled endoscopically, some evidence (very weak, but this is the only pharmacological option) suggests that the use of octreotide and thalidomide reduces the need for iron supplements and/or transfusions (octreotide by decreasing portal flow and thalidomide by suppressing angiogenesis). If the aortic valve disease were stenosis, the option of surgery to improve anaemia should be considered.' (Expert 9, Case 3318-5) (also coded as Codes 1 and 4)

Table 1 (Continued)

Code	Code description	Instances	Experts making at least one such statement (out of 10 experts), *n*	Example statement
9	Participant believes the clinical information provided is not discriminating	3	5	'Because I cannot reach the diagnosis with only the clinical information and an ultrasound. She probably has a gastric tumour, but complementary tests, for example endoscopy and/or echoendoscopy, are necessary to know what kind.' (Expert 6, Case 3330-1) (also coded as Codes 2 and 3)
10	Participant notes that the best course of action depends on variable local resources and/or expertise	2	2	'Presurgical ERCP or laparoscopic cholecystectomy with intraoperative cholangiography could be considered in function of the centre's experience.' (Expert 2, Case 3353-1)

[7%] 'completely confident'). Exact agreement in post-data belief changes between the two time-points occurred in 89 of 150 (59%) instances. The linear weighted kappa value between the two sampled time-points, a statistic that accounts for chance agreement, was only 0.33 (95% confidence interval [CI] 0.21–0.46), suggesting limited consistency. Collapsing responses at each time-point into a 3-point scale (i.e. treating –2 and –1 ratings as equivalent, and +1 and +2 ratings as equivalent) only slightly increased exact agreement (98/150, 65%) and the weighted kappa (0.37, 95% CI 0.25–0.52).

Codes for experts' reflections on their inconsistencies are summarised in Table 2. The most frequent rationale (22 instances) involved experts sharing their current clinical reasoning without explaining why they believed their answer had changed. Several experts indicated a need for additional clinical information (11 instances) or noted that they interpreted the case data differently each time (11 instances). For example, Expert 1, Case 3343-1, stated:

> I gave more importance to the possibility that this patient could have NASH [non-alcoholic steatohepatitis] based on her clinical history.

A few experts indicated that scale point differences were not meaningful to them (e.g. Expert 1: 'I think that [*less likely*] is not so far from [*neither more nor less likely*]'). Occasionally, experts' responses indicated that they had not considered the proposed action suitable in the first place and so were using somewhat artificial reasoning. Expert 10, Case 3350-1, wrote:

> In cases with suspected biliary pancreatitis, a cholecystectomy would be indicated, not [the proposed action]. But seeing that the author for this clinical case proposed an endoscopic ultrasound, I decided to play along. Since they do not want to do MR [magnetic resonance] cholangiopancreatography [in this clinical scenario], I'll give them what they want.

When two or more experts agreed in their post-data belief change (133/150 responses), for 15 (11%) of those responses, experts also agreed in their pre-data beliefs. However, for most responses (89%), pairs or larger groupings gave the same post-data belief change but at least two of each group's experts disagreed in their pre-data beliefs. For example, in Case 3299-2, two experts agreed that the new data had no effect on the suitability of the investigation, but whereas one expert's pre-data belief had been that the investigation was suitable, the other expert had indicated that it was *not* suitable.

Evaluating others' perspectives

After viewing the other experts' responses and rationales, the mean ± standard deviation rating given by experts for how compelling those rationales were was 2.93 ± 0.80 (where 3 = moderately compelling). Experts then switched

Table 2 Codebook: rationales for discrepancies in post-data belief changes over time

Code	Code description	Instances	Experts making at least one such statement (out of 10 experts), *n*	Example statement
1	No explanation (e.g. participant just 'thought aloud' about the case without explaining the discrepancy)	22	9	'The new information makes the proposed treatment less useful, because we need to wait for the histologic results before starting empirical treatment.' (Expert 4, Case 3311-1)
2	Participant wants more information (e.g. laboratory tests)	11	9	'The two options given are to assume (a) that the episode of pain 2 years ago was acute pancreatitis that led to the formation of a pseudocyst that was unrecognised at the time or (b) that she has a cystic tumour. I think we need to try a fine-needle biopsy of the lesion to determine the diagnosis.' (Expert 2, Case 3351-1)
3	Participant changed his or her interpretation of case data (e.g. history, image)	11	6	'Gastric cancer is an indication for *Helicobacter* eradication, but at first we were managing a patient with anaemia. Given the clinical change, cultures and antibiogram should be done to define the antibiotic treatment, but first the cancer should be treated. I think this last reason must be why I changed my answer to B [*less likely*], giving priority to the treatment of the cancer.' (Expert 2, Case 3311-1)
4	Participant believes data given were not discriminating	6	5	'The new information really doesn't add much to what we already knew, except maybe her ulcerative colitis is not active right now, so I think that answer B [*less likely*] is not so far from answer C [*neither more nor less likely*].' (Expert 1, Case 3305-5) (also coded as Code 5)
5	Participant doesn't consider scale anchors to be significantly different	6	3	'In my opinion, A [*much less likely*] and B [*less likely*] are the same. One is more exhaustive than the other. I consider the proposed study useless or of little use: I wouldn't do it. I would do pH-impedance to define whether biliary or alkaline reflux was present.' (Expert 10, Case 3296-3) (also coded as Code 9)
6	Participant believes he or she misinterpreted clinical information the first time	4	3	'I suppose that at that time I thought that 400 UI/mL [*sic*] wasn't enough to consider post-ERCP [endoscopic retrograde cholangiopancreatography] pancreatitis, which continues to be a valid criterion because high amylase levels do not necessarily imply pancreatitis. Re-evaluating the ERCP image, I can now clearly see the opacification of the main pancreatic duct, which persists with a biliary tract cannulated only by the guidewire (without contrast material), and this increases the risk of pancreatitis. A decade ago, this was routine. Nowadays, patients are usually admitted with only a

Table 2 (Continued)

Code	Code description	Instances	Experts making at least one such statement (out of 10 experts), *n*	Example statement
				guidewire and opacification of the main pancreatic duct is avoided. In addition to the poor technique that empties the pancreatic duct badly, with a greater risk of clinical pancreatitis. I misinterpreted these findings.' (Expert 10, Case 3352-5)
7	Participant admits arbitrary change in decision making	4	3	'I cannot explain why I chose a different answer.' (Expert 7, Case 3354-2)
8	Participant learned something that reinforced or changed his or her view since first answering	4	4	'Maybe because I read an article last month related to this topic.' (Expert 7, Case 3318-5) (also coded as Code 12)
9	Participant gave a low post-data belief change answer in the past, when what was intended was to express disagreement with the hypothetical in general (i.e. having a low pre-data belief)	4	3	'I think that there was no indication for pH monitoring before; although there is a peptic stenosis, it is still not indicated. That's why I'm inclined to put choose C (my decision was taken before the new information).' (Expert 7, Case 3296-3)
10	Participant acknowledges confirmation bias	2	1	'I think that having considered a GIST [gastrointestinal stromal tumour] a possible diagnosis, in the earlier round I tried to find elements to support that suspicion. As this test is related to discrepancies, these images changed my view toward excluding GIST.' (Expert 10, Case 3330-1)
11	Participant wants to validate clinical data (e.g. laboratory data or image)	2	2	'Although the CT [computed tomography] report says 2 cm, it looks larger to me.' (Expert 8, Case 3351-1)
12	Participant acknowledges recency effect	2	2	'Some recent complications in patients undergoing colonoscopy might have influenced my decision.' (Expert 6, Case 3296-3)
13	Participant seems to have misread the case	2	2	'I consider that since he has had a cholecystectomy we could do pH monitoring/impedance (for non-acid reflux), and since he doesn't have luminal stenosis, pH monitoring is not contra-indicated.' [Note that the case indicates the patient *does* have luminal stenosis] (Expert 4, Case 3296-3)
14	Participant believes the proposed action is valid, although he or she would consider another proposed action first	2	3	'The patient's personal and family history and her anaemia put her at risk of colorectal cancer. It seems more useful to do a virtual colonoscopy than a barium enema. Of course, if virtual colonoscopy is not available, a barium enema can be done.' (Expert 6, Case 3299-2)

responses 20% of the time. All but one expert changed at least one response across the 15 cases, with a maximum of six responses (40%) changed. In three cases, no experts switched, whereas in one case, five (50%) experts changed their responses.

Experts' general reflections are summarised in Table 3. The most frequent reflection was that there were *multiple partially correct answers* (26 instances), often specifically depending on variable local conditions (eight instances). Expert 1 on Case 3353-1 wrote:

> Both approaches seem sensible. If there is a team that does a high volume of ERCP [endoscopic retrograde cholangiopancreatography], maybe they have the material and skills to resolve the case endoscopically.

Expert 5 on Case 3304-1 stated:

> I strongly agree with [*less likely*]; [*more likely*] also seems acceptable. In any case, in these diseases, treatment is 'trial and error' and we will just have to wait and see the outcome.

Expert 10 on Case 3299-2 stated:

> This case shows one of the limitations of [the SCT]. Both the answers [*more or much more likely*] and [*more or much less likely*] could and can be justified on the basis of external elements (the quality of the first colonoscopy, availability of different equipment – paediatric equipment – quality of the double-contrast barium enema, availability of software for virtual colonoscopy). In other words, in different contexts, the two seemingly opposing positions are valid.

DISCUSSION

We investigated experts' response processes in SCT cases in which expert scoring panel members had opposing views. Findings often contradicted the tested theory of response process that underlies the aggregate scoring methodology, extending the body of research warning against the use of classical SCT methods for assessing examinees.[6–8,10,12,19]

Script concordance test proponents have assumed that the classical SCT approach features enough psychological fidelity to activate well-developed illness scripts. Although SCT methodology requires that cases' proposed actions should all be 'relevant' or 'plausible',[4,5] our findings suggest such actions may not be what experts would ordinarily consider, and thus may not be amongst their automatically triggered set of illness scripts. In other instances, the proposed actions may be amongst experts' ordinary considerations, but experts need more information before evaluating the meaning of new clinical data. Thus, the classical SCT format risks forcing respondents to skip the data collection phase of clinical reasoning prematurely and to move ahead to data interpretation. Finally, experts' narrative comments suggested that differences in responses between experts did not always and exclusively reflect differences of opinion. Rather, experts often were 'filling in the blanks' of cases with local contextual details that differed amongst them, such as those pertaining to the quality of equipment or the proficiency of imaging specialists. Consistent with the notion of illness scripts as complex knowledge structures, perhaps reasoning often requires consideration of the *interaction* of several pieces of data, including such contextual details, rather than only consideration of simple 'action–data' pairs as presented in classical SCTs.

Another major assumption of the classical SCT method is that experts' post-data belief change selections can be interpreted in a straightforward way: specifically, that it is suitable to add up how many experts select each response and then to use those sums to set the scoring key. This study provides empirical evidence for Kreiter's[7] conjecture: that a post-data belief change, such as 'the data make the action *much more* suitable', can only be understood if we know how suitable the subject thought the action was in the first place. In addition, straightforward interpretation is challenged by the inconsistency seen within experts' responses, both before seeing other experts' responses (akin to test–retest reliability) and after. With respect to test–retest reliability, these experts were all highly qualified, held at least moderate confidence in their responses, and only rarely mentioned any lack of knowledge or experience related to the cases and hence, the inconsistency is a somewhat unexpected finding. One possibility is that experts may make different assumptions about important, unspecified case details that vary not only amongst experts but also within individual experts over multiple test occasions. Additionally, experts' responses were not all firmly held; the experts varied in how compelling they found each other's positions to the point of switching answers in several instances. Thus, an expert panel's set of responses on a single test administration, without

Table 3 Codebook: general reflections

Code	Code description	Instances	Experts making at least one such statement (out of 10 experts), *n*	Example statement
1	Participants believe there are multiple partially correct answers	26	8	'The differential diagnosis is between the two entities proposed, and the definitive diagnosis will come from the biopsies. The two arguments are logical and it is important to remember that not all lesions have the "typical" characteristics.' (Expert 2, Case 3330-1) (also coded as Codes 3 and 7)
2	Participant agrees that the case is basically consistent with the proposed action	18	6	'The finding of non-specific erosions does not justify the diagnosis of a relapse; the lack of significant findings on the colonoscopy strongly supports the diagnosis of irritable bowel syndrome.' (Expert 2, Case 3305-5)
3	Participant wants more clinical data (e.g. laboratory tests)	15	8	'The patient needs a more comprehensive workup to confirm the diagnosis. The colonoscopy is not enough, we need the results of laboratory tests and stool tests.' (Expert 8, Case 3305-5)
4	Participant notes that the best course of action depends on variable local resources and/or expertise	8	6	'The availability of diagnostic techniques varies among centres. We should use the best method that is available to us to reach the diagnosis in our patients.' (Expert 6, Case 3299-2)
5	Participant notes some case details are at least partially inconsistent with the proposed action	5	4	'The data are contradictory. A normal US [ultrasound] 2 years ago and a 10-cm lesion now point to a fast-growing lesion. The negative serologic tests for viruses do not rule out the possibility that the patient has liver disease (e.g. fatty liver) that would cause hepatocellular carcinoma.' (Expert 2, Case 3343-1) (also coded as Code 1)
6	Participant considers some others' arguments to be less valid	5	4	'The essence of this system of clinical scenarios is that there is not a single exclusively right answer. Although my choice (C [*neither more nor less likely*]) continues to seem reasonable until better imaging studies can rule out hepatocellular carcinoma, the arguments in B [*less likely*] seem better to me than those in D [*more likely*]. Answer D would imply that an obese person with theoretically moderate alcohol consumption could develop chronic liver disease capable of generating a hepatocellular carcinoma without tumour markers. For all these reasons, based on the experts' opinions (I do not work in this area), I'm inclined to go with B [*less likely*].' (Expert 10, Case 3343-1) (also coded as Codes 1 and 11)
7	Participant thinks aloud about the case in general	5	4	'If a new enteroscopy is done and several lesions are treated, I wouldn't do anything until the patient has clinical manifestations. Thalidomide is useful but not without adverse effects.' (Expert 3, Case 3318-5)

Table 3 (Continued)

Code	Code description	Instances	Experts making at least one such statement (out of 10 experts), *n*	Example statement
8	Participant challenges the appropriateness of actions taken in the case description	4	4	'If a patient had early-stage gastric cancer, I would not treat it with *Helicobacter* eradication. I would get a gastrectomy and thus avoid follow-up for the rest of my life, which would be long for someone diagnosed with early gastric cancer without positive lymph nodes.' (Expert 10, Case 3311-1)
9	Participant recognises he or she misinterpreted or changed the interpretation of case data (e.g. history)	3	2	'I misinterpreted the answer.' (Expert 9, Case 3353-1)
10	Participant wants more specific information; certain case specifics are unnecessarily ambiguous	2	2	'Maybe the first description could have given a little more specific information.' (Expert 1, Case 3330-1)
11	Participant acknowledges his or her own limited knowledge	2	1	'Because of my poor level of expertise in this specific area it is easy for me to change my mind if the experts in favour of B insist that corticosteroids are indicated. The reason I'm inclined to use the suppositories is to give the patient time to improve his condition with a less aggressive treatment without the undesirable systemic effects of corticosteroids.' (Expert 10, Case 3304-1)

subsequent debate, is a rather limited sample of the panel's underlying beliefs.

Limits to the generalisability of the findings in this study include that the study focused on one SCT, within one specialty and nationality, for cases that had previously proven to be controversial. That said, in a new area of research, it is valuable to show that a novel or unexpected finding *can* be observed, even if its generalisation is not yet known,[20] and, as with all theory testing, theory-inconsistent findings should prompt serious reconsideration. Additionally, the fidelity of our cases is likely to be higher than usual as they were adapted from real patients' histories and incorporated rich details and findings. As such, we suspect we are underestimating the extent to which experts have construct-irrelevant differences in interpretations and assumptions about cases in the broader SCT literature and thus, we are actually overestimating reliability. It might be argued that our cases were so rich that they diverged from SCT guidance to be 'brief' and 'ill defined'.[4] However, the details of our cases were sufficiently limited to cause experts to often request additional clinical information. More fundamentally, from a cognitive psychology perspective, the assertion that limited details will best elicit script-like cognitive processing does not seem well supported because scripts can be triggered by the rich, incidental contextual details that accompany suitable situations for activating memories.[21,22]

In considering interesting future directions, we note that SCT research has inadvertently yielded rich data on clinical controversies and practice variations amongst experts. Sometimes, relevant clinical evidence may be available but not widely known amongst clinicians, which suggests educational opportunities. Alternatively, SCTs may identify areas in which clinical evidence truly is lacking, thereby helping to better target future clinical research,

akin to a *value of information analysis*.[23] We encourage the reanalysis of prior SCT work from this perspective.

Finally, controversial cases may be valuable stimuli in an assessment *for* learning that better reflects the complexity of medical decision making.[24] Rather than examinees being told their answers are correct or incorrect based on how many experts agreed with them, they might be asked to view the diverse expert responses, rationales and available clinical evidence and then to engage in debate. In this, the learning objectives might refer not to knowing the 'correct' answer, but to gaining better understanding of all sides of an underlying controversy. Even when examinees do not change their beliefs, knowing the controversies, they might calibrate their *certainty* and apply more divergent thinking going forward. Some studies have already used SCTs as assessments for learning,[25,26] although these have been grounded in the aforementioned problematic SCT scoring practices. The Practicum Script programme from which this research draws[18] was originally based on classical SCT methods and has been modified in consideration of the validity issues addressed here and in related research. Further research evaluating how such SCT-like assessment-for-learning applications stimulate improved reasoning would be valuable.

CONCLUSIONS

This study found complex and previously unmeasured phenomena at play in experts' responses to SCT cases. This suggests the value of bringing new curiosity about how examinees respond to such cases, as well as skepticism about past methods for interpreting and judging the reasoning presumed to underlie their responses. As the field reconsiders the SCT and how the complexity of authentic clinical practice can be represented, we hope our findings support new ways forward that better enhance our understanding and facilitation of reflective, effective clinical reasoning.

Contributors: ML was chiefly responsible for the study design, the coding, analysis and interpretation of data and the drafting of the publication. EH, EP, JM, CB and GB all participated substantively in the study design, coded qualitative data and contributed to data interpretation and the drafting of the paper. All authors (ML, EH, EP, JM, CB and GB) approved the final version of the manuscript.
Acknowledgements: none.
Funding: none.
Conflicts of interest: the Practicum Institute for Applied Research in Health Sciences Education administers the Practicum Script assessment system discussed in this manuscript. ML and GB have no obligating relationships with the Practicum Institute and do not stand to gain or lose financial or other resources as a result of any inferences drawn about the assessment system or the extent to which it is used by others.
Ethical approval: The Institutional Review Board at the University of Illinois at Chicago approved the protocol.

REFERENCES

1 Durning SJ, Lubarsky S, Torre D, Dory V, Holmboe E. Considering 'nonlinearity' across the continuum in medical education assessment: supporting theory, practice, and future research directions. *J Contin Educ Health Prof* 2015;**35** (3):232–43.
2 Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The script concordance test: a tool to assess the reflective clinician. *Teach Learn Med* 2000;**12** (4):189–95.
3 Charlin B, Boshuizen H, Custers EJ, Feltovich PJ. Scripts and clinical reasoning. *Med Educ* 2007;**41** (12):1178–84.
4 Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: from theory to practice: AMEE Guide No. 75. *Med Teach* 2013;**35** (3):184–93.
5 Fournier J, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC Med Inform Decis Mak* 2008;**8** (1):18–24.
6 Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. *Med Educ* 2013;**47** (12):1175–83.
7 Kreiter CD. Commentary: the response process validity of a script concordance test item. *Adv Health Sci Educ Theory Pract* 2012;**17** (1):7–9.
8 Ahmadi S-F, Khoshkish S, Soltani-Arabshahi K, Hafezi-Moghadam P, Zahmatkesh G, Heidari P, Baba-Beigloo D, Baradaran HR, Lotfipour S. Challenging script concordance test reference standard by evidence: do judgments by emergency medicine consultants agree with likelihood ratios? *Int J Emerg Med* 2014;**7** (1):34.
9 Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script concordance testing: a review of published validity evidence. *Med Educ* 2011;**45** (4):329–38.
10 Lineberry M. Missing the mark: the faulty logic of aggregate scoring in script concordance tests. *Med Educ* 2014;**48** (11):1038–40.

11 Lubarsky S, Dory V, Meterissian S, Lambert C, Gagnon R. Examining the effects of gaming and guessing on script concordance test scores. *Perspect Med Educ* 2018;**7** (3):174–81.
12 Wan MS, Tor E, Hudson JN. Improving the validity of script concordance testing by optimising and balancing items. *Med Educ* 2018;**52** (3): 336–46.
13 Power A, Lemay J-F, Cooke S. Justify your answer: the role of written think aloud in script concordance testing. *Teach Learn Med* 2017;**29** (1):59–67.
14 Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995;**50** (9):741–9.
15 Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. *Psychol Rev* 2004;**111** (4):1061–71.
16 Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas* 2013;**50** (1):1–73.
17 Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ Theory Pract* 2014;**19** (2):233–50.
18 Hornos EH, Pleguezuelos EM, Brailovsky CA, Harillo LD, Dory V, Charlin B. The Practicum Script Concordance Test: an online continuing professional development format to foster reflection on clinical practice. *J Contin Educ Health Prof* 2013;**33** (1):59–66.
19 See KC, Tan KL, Lim TK. The script concordance test for clinical reasoning: re-examining its utility and potential weakness. *Med Educ* 2014;**48** (11): 1069–77.
20 Campitelli G. Answering research questions without calculating the mean. *Front Psychol* 2015;**6**:1379.
21 Smith ER, DeCoster J. Dual-process models in social and cognitive psychology: conceptual integration and links to underlying memory systems. *Pers Soc Psychol Rev* 2000;**4** (2):108–31.
22 Strack F, Deutsch R. Reflective and impulsive determinants of social behavior. *Pers Soc Psychol Rev* 2004;**8** (3):220–47.
23 Keisler JM, Collier ZA, Chu E, Sinatra N, Linkov I. Value of information analysis: the state of application. *Environ Syst Decis* 2014;**34** (1):3–23.
24 Dannefer EF. Beyond assessment of learning toward assessment for learning: educating tomorrow's physicians. *Med Teach* 2013;**35** (7):560–3.
25 Fernandez N, Foucault A, Dubé S, Robert D, Lafond C, Vincent A-M, Kassis J, Kazitani D, Charlin B. Learning-by-concordance (LbC): introducing undergraduate students to the complexity and uncertainty of clinical practice. *Can Med Educ J* 2016;**7** (2):e104–13.
26 Foucault A, Dubé S, Fernandez N, Gagnon R, Charlin B. Learning medical professionalism with the online concordance-of-judgment learning tool (CJLT): a pilot study. *Med Teach* 2015;**37** (10):955–60.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article

Appendix S1. Example of Practicum Script case and study questionnaire items (representing one of the actual cases used in the study).

Received 4 May 2018; editorial comments to authors 25 June 2018; accepted for publication 28 December 2018

