BMC Medical Education

RESEARCH Open Access

Script concordance test issues, the trail of expert calibration

Yannick Perdrix[1,2,3*], Nicolas Pinsault[2] and Eric Dionne[1]

Abstract

Background The Script Concordance Test (SCT) is an assessment tool for clinical reasoning that incorporates uncertainty and depends on expert judgment to identify valid responses. Accurate calibration of expert judgment is important for maintaining validity and reliability; however, the literature has rarely addressed this issue and only through statistical methods. This study aimed to compare calibration strategies using statistical moderation and qualitative inspection.

Methods Sixteen experts ($n = 16$) were recruited to complete 21 clinical vignettes, providing justification for each response. Seven calibration strategies—quantitative, qualitative, and mixed—were then analyzed using the Rasch Facet Model, with particular attention to expert homogeneity, data–model fit, and the quality of expert responses.

Results None of the strategies improved expert homogeneity. However, mixed strategies enhanced data–model fit and response quality, and helped address issues related to response process and content validity.

Conclusions Calibrating expert judgment using a mixed strategy appears valuable for improving the quality of expert-generated data within an SCT framework. This calibration may address specific psychometric limitations of SCTs and enhance training quality through Learning by Concordance methods.

Keywords Script concordance, Calibration, Expert, Mixed methods

Background

Clinical reasoning is regarded as the cornerstone of patient care by healthcare professionals. Higgs defines it as "a complex process in which critical analysis and reflection take place in the context of action and interaction with the patient" [1]. This reasoning process, whose primary purpose is decision-making, combines resources—including knowledge [2, 3]—that can be structured into action-oriented knowledge networks, known as scripts [4]. Health sciences education must teach and assess students' clinical reasoning using situations that are often complex, contextualized, and uncertain to reflect clinical reality as accurately as possible [5]. Uncertainty, an essential aspect of clinical reasoning, involves recognizing one's own ignorance [6] and not knowing what to do, think, or feel [7]. These situations present significant challenges because our ability to develop and assess these concepts remains incomplete [8].

Several instruments are available to develop or assess clinical reasoning, with some specifically designed to address uncertainty. For assessment, Daniel's [9] systematic review provides an overview of various approaches to evaluating clinical reasoning, analyzing their validity and feasibility, and highlighting their respective advantages and limitations. These instruments aim to clarify

*Correspondence:
Yannick Perdrix
yannick.perdrix@ies-reunion.fr
[1]Faculty of Education, University of Ottawa, Ottawa, ON, Canada
[2]Laboratory TIMC (UMR CNRS 5525), Faculty of Medicine, University of Grenoble Alpes, Grenoble, France
[3]Centre Hospitalier Universitaire de la Réunion, Saint Pierre, France

students' reasoning through scores that may be dichotomous or polychotomous. In polychotomous scoring, partial credit may be given for partially relevant reasoning, which is possible, for example, with the Script Concordance Test (SCT). The SCT is distinctive because it can be computerized, remote, asynchronous, and scalable; it enables automated grading and requires fewer resources than clinical activities such as Objective Structured Clinical Examinations (OSCEs). Additionally, it allows for the deliberate introduction of uncertainty inherent to clinical practice, making it a more authentic learning activity. This feature makes the SCT a unique tool: resource-efficient, scalable, and capable of simulating real-world uncertainty.

In an SCT [10–12], respondents are presented with a clinical vignette and a hypothetical diagnosis, investigation, or intervention. After receiving additional information, they indicate on a five-point scale (– 2, – 1, 0, + 1, +2) the extent to which this new information affects the hypothesis. A + 2 means the hypothesis is strongly reinforced, + 1 moderately reinforced, 0 neither reinforced nor weakened, – 1 weakened, and – 2 strongly weakened. To identify valid responses, a panel of experts completes the activity, and each response category is weighted based on their judgment. While complete agreement among experts is not expected and some variation is natural, a high level of agreement is necessary. This balance is essential for the SCT: minimal variance suggests little uncertainty and makes the test similar to a multiple-choice question, whereas excessive variance can compromise validity, particularly regarding inference. Expert judgment is crucial, because their responses directly inform the scoring key.

Although widely used in education since the 2010s, several studies have identified major shortcomings in the SCT regarding validity and reliability, particularly because certain variables may influence or bias expert judgment. The main concerns relate to response process validity, such as: (1) experts may find the proposed hypothesis unsuitable and must engage in reasoning they would not normally consider [13], (2) the "0" point on the scale may be interpreted inconsistently or used as a default when experts are unsure how to respond [14], (3) the extreme values (– 2 and + 2) on the response scale are often underused [15], (4) the selected response scale may not accurately reflect the expert's actual reasoning [14–17]. These variables may also affect reliability, specifically when experts identify valid responses and when respondents choose the correct options.

Given these shortcomings, most studies on SCT design have focused only on analyzing the reliability of respondents' scores, often removing or adding items to achieve acceptable reliability (e.g., Cronbach's alpha > 0.80). However, they have not considered the quality of experts' responses, despite these responses forming the basis for calculating respondents' scores. However, there is a strategy that first calibrates experts' judgments before considering respondents' scores [18, 19]. Calibration involves ensuring that experts, under identical conditions, provide the same information when presented with the same questions [19], and that this information remains consistent across experts. To achieve calibration, a moderation process may be used. Crisp [20] identified three types of moderation: statistical, inspection-based, and social. Statistical moderation uses quantitative analysis to adjust scores for differences in expert severity. Inspection-based moderation qualitatively examines experts' scores and adjusts them as needed to address various errors. Social moderation brings together a new panel of experts who iteratively review the information and make adjustments through consensus. However, social moderation is highly demanding and generally impractical in standard SCT development.

Given the structural variance inherent in the SCT, who can be defined as the variation in responses that is naturally present due to the construction of the SCT, and the importance of expert judgment, it is important to investigate the effects of different moderation methods, or their combination, on the calibration of judgment. To the best of our knowledge, only Blais et al. [21]. have examined the moderation process in calibrating expert judgment, relying exclusively on statistical moderation using the Rasch model and focusing on expert severity and agreement. No study has examined the effects of combining statistical and inspection-based methods. To address this gap, this exploratory study aimed to compare statistical and inspection-based moderation strategies for calibrating expert judgment in SCT administration.

Method

Sample

This study was conducted within an entry-level physiotherapy training program focused on the upper quadrant in the musculoskeletal field. Sixteen experts ($n = 16$), each with more than five years of clinical experience, membership in a recognized professional society, and a weekly patient caseload in the field under evaluation, participated. All had completed at least 60 h of additional field-specific training. Participants were trained using a detailed 25-minute video and a test case before completing the test. All participants were informed about the ethical aspects of the study and gave their consent to participate. This study has been performed in accordance with the Declaration of Helsinki and has been approved by the Grenoble Alpes Regional Ethics Committee under number CERGA-Avis-2023-10-A1-2023-5.

Instrument

An SCT consisting of 21 clinical vignettes was developed using the reference methodology [12]. The scenarios addressed three domains: diagnosis, investigation, and intervention. Each scenario included a clinical situation, a hypothesis, additional information, and a five-category response scale. Based on studies that combined SCT with think-aloud techniques to capture clinical reasoning [22, 23], a space was incorporated into the test design for experts to justify their responses, allowing the investigation of qualitative variables for inspection-based moderation.

Data collection

The experts received online training on SCTs, including their operation and the process for justifying answers. They then completed the SCT online, selecting a response for each question from the provided scale and briefly justifying their choice.

The multifaceted Rasch model (MRM)

When calibrating expert judgment through statistical moderation, it is important to consider the main variables, hereafter referred to as "facets," that are likely to influence it. In this study, four facets were selected: clinical reasoning of the experts, area of care (diagnosis, investigation, intervention), item difficulty, and position of each response category. To account for all this information, the Many-Facet Rasch Model (MFRM) [24] was used because it allows integration of all variables. This model also enables estimation of the parameters for each facet on a single measurement scale expressed in logits. The specific MFRM model applied was as follows:

$$P_{nijk} = \frac{e^{B_n - D_i - F_j - T_k}}{1 + e^{B_n - D_i - F_j - T_k}}.$$

In this equation, P_{nijk} represents the probability that subject n will select category k for item i in domain j, B_n is the ability of expert n, D_i is the difficulty of item i, F_j is the difficulty of domain j, and T_k is the threshold for category k (difficulty associated with moving from category K − 1 to k). The analyses were conducted using version 4.2.4 of the Facets software [25]. The Rasch model is frequently used to estimate the psychometric properties of measurement instruments. In calibrating expert or judge assessments, as in Blais et al. [21]., the information examined includes the data's fit to the model, measurement error, person separation index, and parameter targeting on Wright's map. The fit statistics (infit and outfit) identify scores that fit the model less well, indicating cases where an expert's score diverges significantly from the model-predicted score for that expert. The Facets software enables identification of misfitting responses for each data point (e.g., expert 1 on item 4), which can be removed from the model. Through an iterative process, the raw scores can be re-modeled to assess whether the psychometric properties improve. This analytical identification of problematic responses is a major advantage of this model compared with classical test theory, which, for example, provides a single "common" measurement error for all items [26] and does not account for the nonlinearity of raw scores [27].

Process of developing and analyzing the strategies

The development and analysis of the different strategies followed three steps. First, a quantitative analysis identified problematic responses for statistical moderation. Second, a qualitative analysis identified problematic responses for inspection-based moderation. These data then formed the basis for the third step, which involved creating and analyzing the different calibration strategies.

Step 1: quantitative analysis - quantitative identification of problematic responses

The raw data (RAW) were analyzed using Facets software to identify statistically unexpected responses, applying a partial credit parametrization.

Step 2: qualitative analysis - qualitative identification of inconsistent expert responses

This step aimed to identify inconsistencies between an expert's selected answer and its justification through qualitative analysis. The 336 expert responses (21 questions × 16 experts) comprising the RAW data set were analyzed. Identified inconsistencies were classified into three categories (strong, moderate and weak) based on their level of evidence, with criteria established by researcher consensus. A strong level of evidence indicated irrefutable proof of inconsistency between the expert's response and justification, or a clear statement of inability to justify the answer. A moderate level of evidence indicated an inconsistency between the answer and its justification. A weak level of evidence referred to justifications that were incompletely formulated, making them potentially consistent or inconsistent.

Step 3: creation and analysis of the strategies

Creation of the strategies This step involved developing various strategies from the RAW data by excluding specific expert responses identified in steps 1 and 2, following defined rules. Three main strategy categories were proposed. The first category removed responses using only a quantitative method (Quanti), as identified by the quantitative Rasch analysis in step 1. The second category relied solely on inspection from a qualitative perspective (Quali), with the progressive exclusion of inconsistent responses identified in step 2, based on their level of evi-

Table 1 Creation of strategies - Presentation of the excluded data according to the strategies

Excluded data:	RAW	Quanti	Quali-1	Quali-2	Quali-3	Mixed-1	Mixed-2	Mixed-3
Inconsistent responses identified through the quantitative Rasch analysis conducted in step 1		X						
Strong level - inconsistent responses identified through the qualitative analysis conducted in step 2			X	X	X		X	X
Moderate level - inconsistent responses identified through the qualitative analysis conducted in step 2				X	X		X	X
Weak level - inconsistent responses identified through the qualitative analysis conducted in step 2					X			
Inconsistent responses identified by both the quantitative Rasch analysis and the qualitative analysis (all levels)						X	X	X
Inconsistent responses, not identified in step 1 but with a residual greater than 2 logits* identified during the quantitative analysis								X

**A residual greater than 2 logits indicates a substantial discrepancy between the model's expected response and the category selected by the expert, exceeding two response categories. In such cases, the response meaning is effectively reversed (for example, an expected value of + 2 and an actual value of −1)*

Table 2 Indicators and benchmarks used in the quantitative analysis of strategies

Indicators	Expected value
Homogeneity of experts	
Expert separation index	0
Reliability index	0
Number of strata	1
Difference between expert's abilities	0
Data-model fit	
Outfit standardized (OutFit Stzd)	[−2.0; 2.0]
Infit standardized (IntFit Stzd)	[−2.0; 2.0]
Difference between actual and expected correlation PTMEA - PTEXP	0
Quality of expert responses	
Total number of unexpected responses	0
Maximum number of unexpected responses for a single expert	0
Effect of the strategy on the data	
Percentage of responses excluded	0

dence. The third category adopted a mixed perspective, combining statistical and inspection methods to integrate quantitative and qualitative approaches (Mixed). For the qualitative and mixed strategies, several progressive strategies were applied. Table 1 presents the different strategies according to the excluded data.

Analysis of the strategies The comparison of the different strategies used to address the research objective was based on the number of problematic responses identified regarding response process validity (steps 1 and 2), as well as the results of a quantitative analysis using the Rasch model (Table 2).

Quantitative analysis using Rasch modeling was conducted with Facets software to assess the psychometric properties of each strategy. The analysis examined expert homogeneity, data-model fit, the statistical quality of expert responses, and the effect of each strategy on the data. Table 2 presents the indicators and expected values used to compare the strategies.

The homogeneity of the experts was examined using their ability values in logits (unit of measurement in Rasch modeling), which quantified the differences between the highest- and lowest-performing experts, the separation index, the reliability of this separation, and the number of response strata. Unlike typical test respondents, experts are expected to be highly homogeneous. Therefore, a minimal gap between expert abilities, very low separation and reliability, and a number of response strata close to 1 [28] are expected.

The fit of the data to the model was assessed using infit and outfit statistics, with an interval of – 2 to + 2 indicating potential misfit [28–30]. Additionally, the difference between the actual correlation statistic (PTMEA) and the expected correlation (PTMEA-EXP) was expected to be zero or very small, reflecting well-adjusted item–total correlations and non-deviant expert behavior.

The statistical quality of the experts' responses was evaluated by counting unexpected responses, which should be minimal.

The effect of strategies on the data referred to the number of excluded data points. A strategy that excluded many data points would lose relevance; therefore, a null or very small number of excluded data points was expected.

Results

Results of steps 1 and 2

Of the 336 data points, quantitative analysis of the initial (RAW) data identified 11 unexpected responses (3.0%) across the entire data set. These responses were distributed across nine items, with a maximum of two unexpected responses for items 10 and 16. Eight experts were involved; seven had one unexpected response each, and one expert (No. 12) had four unexpected responses.

The qualitative analysis in step 2 identified 37 inconsistencies, accounting for 11% of the data. Among these, 6 responses were classified as strong evidence, 11 as moderate, and 20 as weak. Table 3 presents the results.

Among the 11 unexpected responses identified by the quantitative analysis, only three were also identified by the qualitative analysis. Thus, eight responses were considered problematic by the quantitative analysis but showed no inconsistencies in the qualitative analysis. Additionally, the quantitative analysis did not identify any inconsistencies classified as having a strong level of evidence in the qualitative analysis. The implementation of qualitative analysis therefore provides added value by detecting problematic responses that the quantitative analysis does not identify. The Quali and Mixed strategies are more effective in detecting responses with issues affecting the validity of the cognitive process.

Results of step 3

The initial results showed that one expert responded more inconsistently than the others. Specifically, for the RAW, Quali-1, and Mixed-2 analyses, Expert 12 had standardized Outfit and Infit adjustment values (Stzd) greater than 3, as well as four unexpected responses. A detailed review of Expert 12's response pattern was conducted. This analysis indicated that Expert 12's choices differed significantly from those of the other experts for five vignettes, including the four identified in the quantitative analysis. Although the justification initially appeared consistent with the selected answers, the working hypothesis was that Expert 12 might hold erroneous conceptions regarding these specific questions. An in-depth analysis based on the theoretical and clinical foundations of physiotherapy was conducted independently and blindly by two researchers (YP and ML). Both researchers judged the responses to be invalid, with complete agreement. Consequently, the results of Expert 12 for these questions were excluded from the Mixed-2 strategy, which was then re-analyzed quantitatively. These data were not excluded from the RAW data set to preserve its original state, nor from the Quali-1 strategy, which cannot benefit from quantitative analysis. At this stage, only the Mixed strategies appear capable of detecting responses with content validity issues. The results of the analyses of the different expert calibration strategies are presented in Table 4.

Homogeneity

Regarding the homogeneity of the RAW scores, the expert separation index was 0.44, the reliability index was very low (0.16), the number of strata was close to 1.00, and the expert scores ($n = 16$) ranged from minimum to maximum within 0.93 logit. These values were within the expected range.

For the other calibration strategies, the results show that the Quanti strategy significantly changes expert homogeneity: the separation index increased from 0.44 to 0.99, the reliability index rose from 0.16 to 0.50, the number of strata increased from 0.92 to 1.65, and the score range widened by approximately 50% (from 0.93 to 1.51). For the Quali and Mixed strategies, the data are relatively similar, with homogeneity values close to or slightly higher than those of the RAW strategy. However, the Quali-2 strategy produced values very similar to the RAW strategy.

Data-model fit

Across all strategies, results show minimal differences in both Outfit and Infit indices, indicating that the MFMR Rasch model was appropriate for modeling scores for all strategies, despite the small sample size ($n = 16$). The logit difference between the actual item–total correlation (PTMA) and the expected value (PTMEX) should approach zero if the data and model are well aligned. For the RAW strategy, only eight correlation values had differences smaller than 0.10. The Quali strategies produced broadly similar results. In contrast, the Quanti and Mixed strategies performed better, with 11–14 values showing differences smaller than 0.10.

Table 3 Results of the quantitative and qualitative analyses in steps 1 and 2

		Quantitative Analysis		Subtotal	Total
		Unexpected	Expected		
Qualitative Analysis	Inconsistencies - Strong evidence	0	6	6	37
	Inconsistencies - Moderate evidence	1	10	11	
	Inconsistencies - Weak evidence	2	18	20	
	Consistent responses	8	291	299	
Total		11	325	336	

Table 4 Results of the analyses of the different expert calibration strategies

Indicators	RAW	Quanti	Quali-1	Quali-2	Quali-3	Mixed-1	Mixed-2	Mixed-3
Homogeneity of experts								
Expert separation index	0.44	0.99	0.71	0.42	0.53	0.67	0.65	0.59
Reliability index	0.16	0.50	0.29	0.15	0.22	0.31	0.30	0.26
Number of strata	0.92	1.65	1.28	0.89	1.05	1.23	1.20	1.11
Difference between expert's abilities	0.93	1.51	1.26	1.00	1.18	1.04	1.16	1.28
Data-model fit								
OutFit Stzd between – 1.9 and 1.9	14	16	15	14	13	14	16	16
IntFit Stzd between – 1.9 and 1.9	14	15	15	14	13	12	16	16
PTMEA - PTExp between 0 and 0.09	8	12	10	9	9	11	12	14
PTMEA - PTExp between 0.1 and 0.19	7	4	6	6	5	4	3	2
PTMEA - PTExp greater than 0.2	1	0	0	1	2	1	1	0
Quality of expert responses								
Percentage (N) of unexpected responses	3 (11)	1 (5)	3 (9)	3 (10)	3 (9)	3 (11)	3 (11)	3 (9)
Maximum number of unexpected responses for a single expert	4	2	1	4	4	4	2	2
Effect of the strategy on the data								
Percentage (N) of responses excluded	0 (0)	3 (11)	3 (11)	5 (17)	11 (37)	1 (3)	4 (14)	4 (15)

N Number, *Stzd* Standardized, *PTMEA* Actual correlation, *PTMEX* Expected correlation, *max* Maximum, *min* Minimum

Expert's response quality

The number of unexpected responses ranged from 5 (1.5%) to 11 (3%), which was relatively low compared to the total number of expert decisions. The RAW data showed 11 unexpected responses, with a maximum of 4 for a single expert, accounting for 25% of that expert's responses. The Quanti strategy significantly improved the total number of unexpected responses compared to the RAW data, while other strategies had no effect or only a slight favorable effect. For the maximum number of unexpected responses per expert, some Quali or Mixed strategies had no effect compared to RAW, whereas others reduced this number by half, improving response quality from a statistical point of view.

Effect of the strategies on the data

Results show that five strategies (Quanti, Quali-1, Quali-2, Mixed-2, Mixed-3) excluded a similar proportion of responses, ranging from 3.0% to 5.0%. Mixed-1 excluded the fewest responses (1%), while Quali-3 excluded the most (11.0%).

The results of this study show that mixed strategies appear to be slightly better than the RAW, Quanti, and Quali strategies. This advantage is not mainly due to improved expert homogeneity, which remains similar across strategies, but results from the combination of other factors, such as data–model fit, the quality of expert responses, and the detection of problematic responses related to response process validity and content validity.

Discussion

In this study, the RAW data indicated good overall expert homogeneity. None of the developed strategies improved homogeneity. This contrasts with Blais et al. [28], who found that calibrating experts (n = 50) and excluding certain data improved homogeneity. Notably, our RAW data (separation = 0.44, reliability = 0.16) were significantly better than those reported by Blais (separation = [1.45–1.63], reliability = [0.68–0.73]), and also better than Blais's best calibration strategy (separation = 0.56, reliability = 0.24). We hypothesize that our expert sample was initially very homogeneous, likely because of our recruitment criteria, resulting in a floor effect that limited further improvement in homogeneity.

Conversely, the RAW data showed weaknesses in fit statistics and expert response quality. Different calibration strategies improved these parameters, significantly in some cases (Mixed-2 and Mixed-3). Therefore, the effect of data calibration strategies appears to be primarily in enhancing response quality and data–model fit, rather than increasing expert homogeneity.

Qualitative analysis was conducted during Step 2 to assess response process validity and again at the beginning of Step 3, although this was not originally planned in the methodology. Analysis of unexpected responses to the strategies revealed a content validity issue for one expert and identified five additional inconsistent responses. This demonstrates a notable complementarity between the two approaches, which benefits mixed strategies. Quantitative analysis can *detect* problematic responses, while qualitative analysis can identify the nature of these problems.

The Quanti strategy relied on statistical modeling of RAW data with the Rasch Multi-Facet Model. It identified only 8% of the inconsistencies found in the qualitative analysis and failed to detect any high-level inconsistencies. Additionally, the eight unexpected responses missed by the qualitative analysis did not compromise

the validity of the response process or content, and would have been incorrectly excluded. This highlights a major limitation of using quantitative strategies alone and demonstrates a disconnect between mathematically based quantitative approaches and content- or process-based qualitative approaches. Moreover, the Quanti strategy significantly reduced experts' performance indicators, supporting the need for caution when applying quantitative calibration methods.

To the best of our knowledge, this is the first study to use a mixed-method approach to calibrate expert judgments. Based on our findings, several calibration strategies can be considered following expert responses to an SCT:

1) No calibration: no calibration may be possible if experts are consistent due to strict recruitment criteria; however, this consistency can only be confirmed after calibration analysis. Additionally, issues with content validity and response processes are likely, so this strategy should be avoided.
2) Statistical calibration only: statistical calibration alone can improve the quality of expert responses and data–model fit, but it may reduce consistency. Problems with content validity and response processes may still occur.
3) Inspection-based calibration only: inspection-based calibration alone may be feasible if experts are sufficiently homogeneous and could improve the validity of the response process, but it does not fully address content validity issues.
4) Implementing a mixed calibration that combines statistical and inspection-based approaches offers several advantages. The statistical component enables verification and, if necessary, correction of expert homogeneity, as in the Blais study, and helps identify unexpected expert responses or target content validity analysis. The qualitative component allows identification and exclusion of responses with validity issues in the response process, as well as targeted content validity analysis based on responses flagged by the quantitative step. Although a full content review of all responses is possible, it would require greater resources.

We recommend mixed calibration, which combines statistical and qualitative inspection, as essential in SCT contexts. This approach addresses several key limitations reported in the literature and provides more reliable and valid data for student scores. Although mixed calibration requires slightly more resources than other methods, it would seem to offer a favorable cost–benefit balance regarding accuracy and validity.

In this study, the number of experts was selected to achieve minimum statistical power while remaining small enough to allow for qualitative data processing within the available resources.

Although this study focuses on SCT, its findings may also be relevant to Learning by Concordance (LbC). Combining quantitative analysis to identify problematic responses with targeted analysis of content validity appears to be a promising approach to improving LbC quality, which depends heavily on response content.

Future research is important to confirm these findings, particularly regarding the floor effect on homogeneity, and to examine related perspectives, such as the effect of calibration on the reliability of respondents' scores, qualitative analysis of the nature of experts' errors, and the value of using artificial intelligence to automate qualitative analysis. Future studies should also investigate this type of analysis in a more heterogeneous expert panel. Additionally, studying the effect of social moderation would help objectively assess the cost–benefit balance and move beyond assumptions.

Limitations

These results are based on a relatively small sample, with 21 items and 16 experts, and are limited to the specific disciplinary context of upper quadrant musculoskeletal physiotherapy. Moreover, the expert panel in this study is somewhat atypical, as it is highly homogeneous, in contrast to what is typically observed in SCTs. Consequently, caution is warranted when considering the transferability and applicability of the findings to more heterogeneous panels.

Conclusion

This study examined expert judgment calibration in SCTs by introducing qualitative variables. The results show that mixed calibration achieves the best calibration when both quantitative and qualitative indicators are considered. The use of a Rasch Multi-Facet Model was relevant for achieving high analytical precision. This study offers a new perspective for addressing weaknesses identified in the literature and may contribute to improving both SCTs and LbC in health professions education. However, further studies are needed to confirm these results.

Abbreviations

LbC	Learning by Concordance
Max	Maximum
MFRM	Many-Facet Rasch Model
MRM	The Multifaceted Rasch Model
Min	Minimum
N	Number
OSCE	Objective Structured Clinical Examinations
PTMEA	Actual correlation
PTMEX	Expected correlation
SCT	Script Concordance Test
Stzd	Standardized

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-026-08732-8.

Supplementary Material 1.

Acknowledgements
We sincerely thank Mathieu Lothore (ML) for his valuable contribution to the qualitative analysis.

Authors' contributions
YP: Conceptualization, Methodology, Data collection, Data analysis, Writing—Original Draft and review, Visualization, Project administration NP: Data collection, Review and Editing, Supervision, Project Administration ED: Conceptualization, Methodology, Data analysis, Validation, Writing, Review and Editing, Supervision.

Funding
This work was supported by the Fondation APICIL (grant no. 2002.24) and the Association Nationale pour la Formation des Hospitaliers Océan Indien (ANFH-OI; grant no. 60108).

Data availability
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This study has been performed in accordance with the Declaration of Helsinki and has been approved by the Grenoble Alpes Regional Ethics Committee under number CERGA-Avis-2023-10-A1-2023-5.
During recruitment and in accordance with the procedure approved by the ethics committee, participants were informed of the purpose of the study, the data collected, its anonymization, their storage location, their duration of use, and the use that could be made of them. Their free and informed consent was then obtained, with no consequences in the event of refusal. Only participants who gave their consent were included.

Consent for publication
Not applicable.

Competing interest
The authors declare no competing interests.

Received: 5 November 2025 / Accepted: 28 January 2026
Published online: 03 February 2026

References

1. Higgs J, Jensen GM, Loftus S, Trede FV, Grace S. Clinical Reasoning in the Health Professions. 5th éd. Elsevier; 2025.
2. Durning SJ, Jung E, Kim D-H, Lee Y-M. Teaching clinical reasoning: principles from the literature to help improve instruction from the classroom to the bedside. Korean J Med Educ. 2024;36:145–55. https://doi.org/10.3946/kjme.2024.292.
3. Higgs J. Clinical reasoning in the health professions. 3rd ed. Amsterdam: Elsevier/BH; 2008.
4. Charlin B, Boshuizen HPA, Custers EJ, Feltovich PJ. Scripts and clinical reasoning. Med Educ. 2007;41:1178–84. https://doi.org/10.1111/j.1365-2923.2007.02924.x.
5. Aniort J, Trefond J, Tanguy G, Bataille S, Burtey S, Pereira B, et al. Impact of reference panel composition on scores of script concordance test assessing basic nephrology knowledge in undergraduate medical education. Med Teach. 2024;46:110–6. https://doi.org/10.1080/0142159X.2023.2239441.
6. Han PKJ. Uncertainty in medicine: a framework for tolerance. New York, NY: Oxford University Press; 2021.
7. Lee C, Hall K, Anakin M, Pinnock R. Towards a new understanding of uncertainty in medical education. J Eval Clin Pract. 2021;27:1194–204. https://doi.org/10.1111/jep.13503.
8. Dionne E, Grondin J, Latreille M-E. Exploration des scores à un test de concordance de script sous la loupe de la modèlisation de Rasch. In: Mes Éval Compétences En Éducation Médicale Regards Actuels Prospect. Quebec: Presses de l'Université du Quebec; 2017. p. 77–110.
9. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Acad Med. 2019;94:902–12. https://doi.org/10.1097/ACM.0000000000002618.
10. Charlin B, Deschênes M-F, Fernandez N. Learning by concordance (LbC) to develop professional reasoning skills: AMEE guide no. 141. Med Teach. 2021. https://doi.org/10.1080/0142159X.2021.1900554.
11. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review: construction and implementation of script concordance tests. Med Educ. 2012;46:552–63. https://doi.org/10.1111/j.1365-2923.2011.04211.x.
12. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: from theory to practice: AMEE guide 75. Med Teach. 2013;35:184–93. https://doi.org/10.3109/0142159X.2013.760036.
13. Lineberry M, Hornos E, Pleguezuelos E, Mella J, Brailovsky C, Bordage G. Experts' responses in script concordance tests: a response process validity investigation. Med Educ. 2019;53:710–22. https://doi.org/10.1111/medu.13814.
14. Gawad N, Wood TJ, Cowley L, Raiche I. The cognitive process of test takers when using the script concordance test rating scale. Med Educ. 2020;54:337–47. https://doi.org/10.1111/medu.14056.
15. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47:1175–83. https://doi.org/10.1111/medu.12283.
16. Gawad N, Wood TJ, Cowley L, Raiche I. How do cognitive processes influence script concordance test responses? Med Educ. 2021;55:354–64. https://doi.org/10.1111/medu.14416.
17. Power A, Lemay J-F, Cooke S. Justify your answer: the role of written think aloud in script concordance testing. Teach Learn Med. 2017;29:59–67. https://doi.org/10.1080/10401334.2016.1217778.
18. Hardesty DM, Bearden WO. The use of expert judges in scale development: implications for improving face validity of measures of unobservable constructs. J Bus Res. 2004;57:98–107. https://doi.org/10.1016/S0148-2963(01)00295-8.
19. Yan X, Chuang P-L, SAGE Publications Ltd. How do raters learn to rate? Many-facet Rasch modeling of rater performance over the course of a rater certification program. Lang Test. 2023;40:153–79. https://doi.org/10.1177/02655322221074913.
20. Crisp V. The judgement processes involved in the moderation of teacher-assessed projects. Oxf Rev Educ Routledge. 2017;43:19–37. https://doi.org/10.1080/03054985.2016.1232245.
21. Blais J-G, Charlin B, Grondin J, Loye N, Gagnon R. Estimation du degré d'accord entre experts lors du calibrage d'un test.
22. Tedesco-Schneck M. Use of script concordance activity with the think-aloud approach to foster clinical reasoning in nursing students. Nurse Educ. 2019. https://doi.org/10.1097/NNE.0000000000000626.
23. Wan MSH, Tor E, Hudson JN. Examining response process validity of script concordance testing: a think-aloud approach. Int J Med Educ. 2020;11:127–35. https://doi.org/10.5116/ijme.5eb6.7be2.
24. Linacre JM. Many-faceted Rasch measurement [Internet]. 2nd éd. John Michael Linacre Edition; 1994.
25. Linacre JM. Facets Many-Facet Rasch Measurement Computer Program [Internet]. 2025. www.Winsteps.com.
26. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. Mason: Cengage Learning; 2008.
27. Bond T, Yan Z, Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 4e éd. New York: Routledge; 2020. https://doi.org/10.4324/9780429030499.
28. Blais J-G, Charlin B, Grondin J, Loye N, Gagnon R. Estimation du degré d'accord entre des experts lors du calibrage d'un test de concordance de script avec le modèle à facettes de Rasch. In Québec: Presses de l'Université du Québec; 2011.
29. Dionne E. Appliquer Le modèle de Rasch - Défis et pistes de solution. Quebec: Presses de l'Université du Quebec; 2023.
30. Linacre JM. What do infit and outfit, mean-square and standardized mean? Rasch Meas Trans. 2002;16:878–9.

Publisher's Note

Yannick Perdrix is a PhD candidate in a joint doctoral program between the University of Ottawa and Université de Grenoble Alpes. His research focuses on medical education, with particular emphasis on assessing clinical reasoning under uncertainty and metacognition. He also lectures at the School of Physiotherapy in Réunion, France.

Nicolas Pinsault is a physiotherapist and full professor in Grenoble-Alpes University. His research examines contextual effects (placebo/nocebo)—their mechanisms, ethics, and clinical applications (open-label placebo, clinician–patient communication, and care rituals)—with particular attention to how uncertainty shapes clinical decision-making. In parallel, he evaluates primary-care models (direct access; triage for shoulder and low back pain) through pragmatic trials in real-world ambulatory settings, coupled with health-economic analyses. He leads the ThEMAS team within the TIMC laboratory (UMR 5525) and serves as lead or co-lead on several national projects.

Eric Dionne is a full professor in the Faculties of Education and Medicine of the University of Ottawa. He also holds the UOttawa-ISM (*Institut du Savoir Montfort*) Research Chair in Medical Education. His research focuses on the development and validation of data collection tools for measuring or observing simple and complex learning processes. He is also interested in the modeling of test scores and knowledge transfer in the field of measurement.