Modeling Diagnostic Expertise in Cases of Irreducible Uncertainty: The Decision-Aligned Response Model

Martin V. Pusic, MD, PhD, David A. Cook, MD, MHPE, Julie L. Friedman, MD, Jeffrey D. Lorin, MD, Barry P. Rosenzweig, MD, Calvin K.W. Tong, MD, Silas Smith, MD, Matthew Lineberry, PhD, and Rose Hatala, MD, MSc

Abstract

Purpose

Assessing expertise using psychometric models usually yields a measure of ability that is difficult to generalize to the complexity of diagnoses in clinical practice. However, using an item response modeling framework, it is possible to create a decision-aligned response model that captures a clinician's decision-making behavior on a continuous scale that fully represents competing diagnostic possibilities. In this proof-of-concept study, the authors demonstrate the necessary statistical conceptualization of this model using a specific electrocardiogram (ECG) example.

Method

The authors collected a range of ECGs with elevated ST segments due to either ST-elevation myocardial infarction (STEMI) or pericarditis. Based on pilot data, 20 ECGs were chosen to represent a continuum from "definitely STEMI" to "definitely pericarditis," including intermediate cases in which the diagnosis was intentionally unclear. Emergency medicine and cardiology physicians rated these ECGs on a 5-point scale ("definitely STEMI" to "definitely pericarditis"). The authors analyzed these ratings using a graded response model showing the degree to which each participant could separate the ECGs along the diagnostic continuum. The authors compared these metrics with the discharge diagnoses noted on chart review.

Results

Thirty-seven participants rated the ECGs. As desired, the ECGs represented a range of phenotypes, including cases where participants were uncertain in their diagnosis. The response model showed that participants varied both in their propensity to diagnose one condition over another and in where they placed the thresholds between the 5 diagnostic categories. The most capable participants were able to meaningfully use all categories, with precise thresholds between categories.

Conclusions

The authors present a decision-aligned response model that demonstrates the confusability of a particular ECG and the skill with which a clinician can distinguish 2 diagnoses along a continuum of confusability. These results have broad implications for testing and for learning to manage uncertainty in diagnosis.

Assessing clinical expertise is difficult to do well. Considerable psychometric technology has been developed, where items such as cases, questions, or radiographs are administered to clinicians to probe their clinical expertise, with the resulting data scored using a mathematical model to rank performance. Even for some of these well-established formats, such as high-stakes multiple-choice licensure examinations, there is a variable relationship between the test score and clinical competence.[1] We know that a higher examination score is better, but what does a score of x mean when the test taker returns to clinical practice? Furthermore, testing usually operates under conditions that value the certainty of the answers; however, expertise in clinical practice is perhaps best characterized by how a clinician handles uncertainty.[2] Developing testing approaches that embrace uncertainty could be an important step forward for medical educators.

In this report, we will argue that the psychometric models used in traditional item response modeling (IRM) can also be used to more directly model clinical decision making, taking into account the attendant uncertainty.[3–5] In a traditional IRM process, a series of items are presented to a group of individuals with varying ability. The resulting persons-by-items matrix is then used to mathematically derive both the level of difficulty of each item and an overall numerical estimate of the ability of each person.[6] The typical IRM score assigned to a person is the calculated latent "difficulty" of the items for which this person would have a 50% chance of answering correctly. However, this score is a theoretical construct that is disconnected from clinical care. An additional problem with this difficulty scale lies at the expert end where the items may be difficult not because of a knowledge/skill deficiency on the part of the clinician but because there exists irreducible uncertainty around the correct diagnosis, which requires maximal expertise to properly weight.

Consider the specific example we will use in this report—electrocardiogram (ECG) interpretation. The traditional

Please see the end of this article for information about the authors.

Correspondence should be addressed to Martin V. Pusic, Division of Pediatric Emergency Medicine, Boston Children's Hospital, 300 Longwood Ave., CH3311, Boston, MA 02116; email: martin.pusic@childrens.harvard.edu; Twitter: @mpusic.

Acad Med. 2023;98:88–97.
First published online August 9, 2022
doi: 10.1097/ACM.0000000000004918
Supplemental digital content for this article is available at http://links.lww.com/ACADMED/B322.

test assessing skill in ECG interpretation consists of cases representing possible ECG diagnoses, selected for their range of difficulty. Medical students would be tested with items at the easy end of the difficulty scale, cardiology residents using items at the other end. Each ECG case contributes to the numerical score based on the range of difficulty it represents (e.g., easy ECG cases do not contribute much to the assessment of advanced cardiology fellows). Cases where a clear-cut diagnosis is not possible are usually not used since they are thought to be unfair to test takers striving for a "correct" answer.[7] The construct of difficulty, on which traditional IRM is based, does not transparently distinguish between the difficulty of a case due to the test taker's lack of sufficient training and the irreducible difficulty due to the true clinical uncertainty inherent in some ECGs wherein 2 fully capable cardiologists make different diagnoses.

While difficulty as conceptualized in the usual application of IRM translates only indirectly into the realities of clinical practice, the IRM machinery can, with a relatively straightforward reconceptualization, produce metrics that are directly applicable to decisions made in clinical practice, including the uncertain ones. We will term this reconceptualization decision-aligned response modeling (DA-RM), defined as a subset of IRM specifically applied to clinical decision making. The key difference between typical IRM and DA-RM lies in what is being modeled. Typical IRM answers the question: How likely is this person to answer this item correctly? The mid-point of the scale is defined by an item where a person of average ability is 50% likely to answer correctly. By contrast, DA-RM answers the question: How certain is this person of their diagnosis over the alternative? Here the mid-point of the scale is defined by an item where a person of average bias would be equally likely to choose either of the 2 diagnoses. As we will demonstrate, DA-RM is clinically applicable because the scale can be used to quantify a clinician's diagnostic preference on a given item, including their overall sensitivity/specificity tendency (see Table 1 for a comparison of the features of traditional IRM and the proposed DA-RM).

Here, we describe a proof-of-concept study, using a tightly controlled experimental design focused on ECG

Table 1

Conceptual Differences Between Traditional Item Response Modeling (IRM) and Decision-Aligned Response Modeling (DA-RM)

Feature	IRM	DA-RM	Comment
Latent scale being modeled	Probability of answering an item correctly, from low to high.	Probability of deciding on one diagnosis over another, from always choosing diagnosis 1 to always choosing diagnosis 2.	IRM uses intermediate latent constructs (difficulty/ability) based on correctness of the diagnosis; DA-RM uses latent constructs based on the diagnosis.
Basis of test assembly	Selection of items across a range of difficulties, from easy to hard.	Selection of items based on the degree to which they resemble diagnosis 1, diagnosis 2, or are inherently confusable with one another.	Confusable items may not be suitable for IRM since it may be difficult to determine the correct answer; such items are especially valuable for DA-RM.
Individual person metric	Ability, defined as the level of difficulty of the item for which the person is predicted to have a 50% likelihood of being correct.	Bias, defined as the location on the scale where the person is predicted to have a 50% likelihood of choosing either diagnosis 1 or diagnosis 2.	These notions are complementary. A high-ability individual may have a bias tilted in either direction.
Individual item/case metric	Difficulty, defined in terms of an average person's ability to answer the item correctly.	Confusability, defined in terms of an average person preferring diagnosis 1 over diagnosis 2. Items at the center of the scale are maximally confusable.	IRM depends on the ability of the instrument developer to determine a gold standard correct answer for each item; DA-RM does not.
For a person, the mid-point of the scale represents…	… a person who classifies an item/case of average difficulty with 50% accuracy.	… a person of average bias, having no systematic preference for one diagnosis over the other.	Clinical applicability of IRM is limited for individuals because the difficulty/ability scale poorly represents how a clinician would rate uncertain cases. DA-RM is clinically applicable because the scale corresponds to the sensitivity/specificity trade-off made by the individual and can be used to predict that person's diagnosis on a given item.
For an item, the mid-point of the scale represents…	… an item for which an average person would be predicted to answer correctly 50% of the time.	…an item for which a person of average bias would be equally likely to choose diagnosis 1 or diagnosis 2.	The mid-point of the IRM scale for items has limited clinical applicability. The mid-point of the DA-RM scale denotes items of maximum confusability, allowing modeling of (appropriate) uncertainty.
Mathematical model	Multiple options, including graded response model.	Multiple options, including graded response model.	While the mathematical formula may be the same for both models, what is being modeled is different: IRM models the probability of being correct, while DA-RM models the probability of choosing diagnosis 1 over diagnosis 2.
Applicability of construct to clinical practice	The individual with high IRM ability is able to diagnose a wide range of items correctly; their ability to diagnose uncertain (confusable) cases may be underspecified.	The individual with high DA-RM ability is able to assign a given item the right level of diagnostic certainty on a scale between 2 competing diagnostic possibilities; their ability to diagnose more general situations may be underspecified.	These approaches are likely to be complementary, DA-RM operating well when 2 competing diagnoses can be specified and IRM operating well when the situation has multiple competing diagnoses.

interpretation, to demonstrate the statistical approach and potential advantages of DA-RM. Developing psychometric approaches that better represent uncertainty and model expertise will help us maximize our understanding of performance and leverage this understanding for learning.

Method

We present a prospective cohort study of clinicians at different training levels diagnosing ECG cases chosen because they could represent, to a purposefully varying degree, 2 diagnoses that would prompt different management decisions: pericarditis vs ST-elevation myocardial infarction (STEMI).

DA-RM vs IRM

In Figure 1, we demonstrate the DA-RM reconceptualization using 4 example ECG diagnoses that might be the basis for testing ECG skill. In a traditional IRM context, items would be chosen to represent each of the 4 diagnoses (normal, atrial fibrillation, STEMI, pericarditis) that are part of the difficulty construct; cases where the diagnosis might not be clear would be excluded. However, one of the key reasons why pericarditis is difficult to diagnose is that it shares many features of STEMI to the point where some ECGs are not classifiable between the 2 diagnoses. Determining the limits of diagnostic reasoning in confusable cases such as these, and identifying clinicians' varying

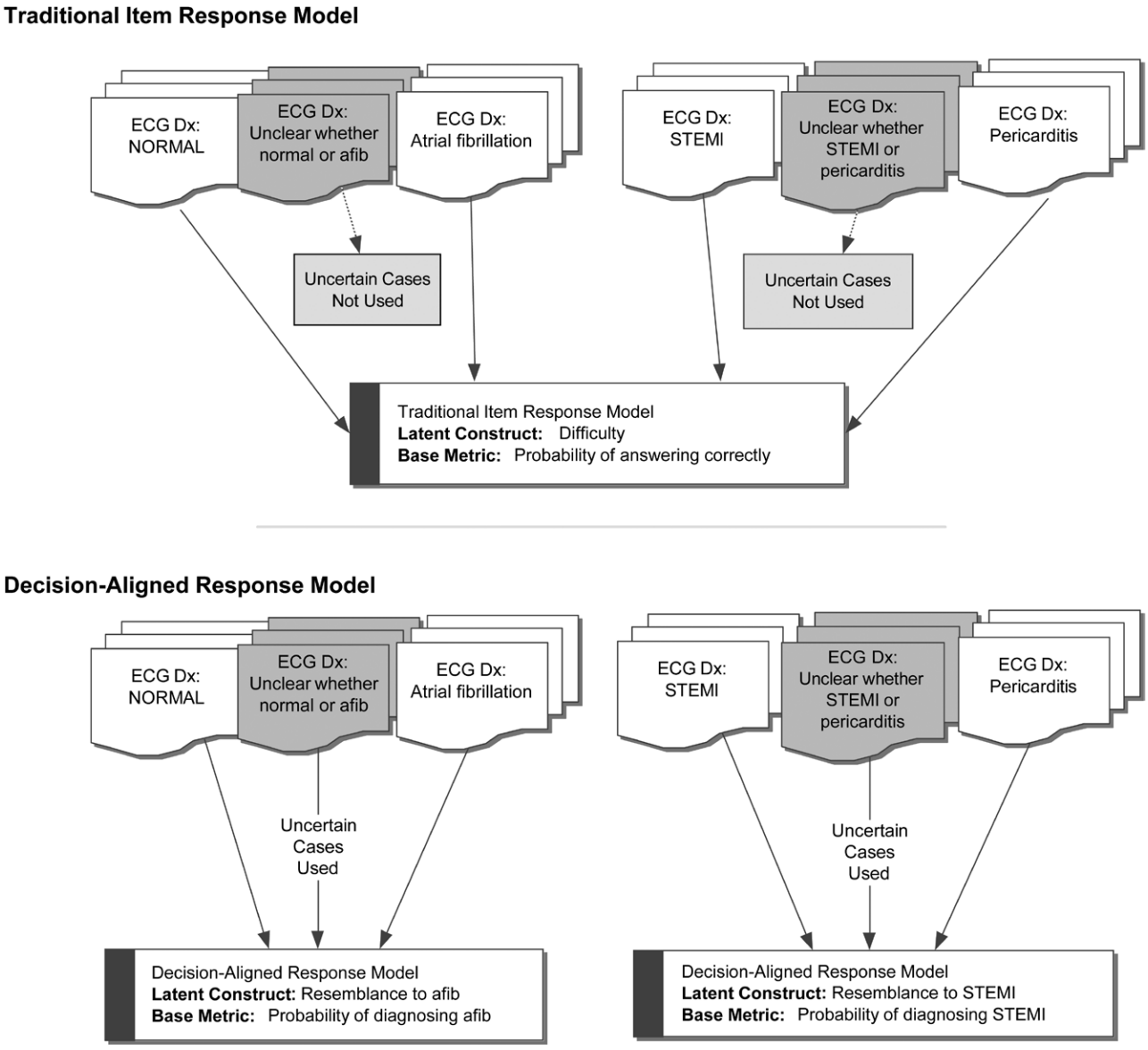

Figure 1 Conceptual framework for traditional item response modeling (IRM) vs decision-aligned response modeling (DA-RM). In traditional IRM (upper panel), items are chosen to represent a latent scale of difficulty ranging from easy to hard, based on the probability of an individual responding to the item correctly. Items are ranked as difficult if they have a low probability of being answered correctly. Uncertain items are systematically excluded. The diagnosis (e.g., pericarditis vs STEMI) influences the scale only indirectly through its degree of difficulty. In DA-RM (lower panel), items are chosen to represent a latent continuous scale between 2 diagnostic poles (e.g., pericarditis or STEMI); items range from prototypical cases at either end of the scale to ones in the middle where an expert clinician might think the item is equally likely to be one diagnosis or the other. Items are ranked by the probability of endorsement of one diagnosis over the other. Thus, the actual diagnosis influences the scale directly. See Table 1 for more about the conceptual differences between these models. Abbreviations: ECG, electrocardiogram; Dx, diagnosis; STEMI, ST-elevation myocardial infarction.

latent abilities to accurately discriminate STEMI from pericarditis, constitutes a potentially useful target for assessments of clinical expertise.

Under the DA-RM reconceptualization, items are chosen by the degree to which they represent either STEMI or pericarditis along a continuum ranging from the poles at either end, defined by clear-cut textbook cases of each diagnosis, to those cases in the middle where even an expert cardiologist would remain appropriately uncertain. The result is a quantifiable test of a clinician's ability to appropriately make the classification decision, pericarditis vs STEMI, along the full continuum of confusability (see Table 1).

Selection of clinical exemplars (diagnoses)

We investigated the following clinical decision: Does this ECG demonstrate either pericarditis or STEMI? We selected these diagnoses because their ECG features were confusable and the resulting clinical management varied substantially. ECGs with STEMI features include localized ST-segment elevation in the ECG leads corresponding to the ischemic anatomical region, with reciprocal ST-segment depression in the ECG leads relating to the nonischemic region. ECGs with pericarditis features include diffuse ST-segment elevation except in lead aVR where ST-segment depression and PR-segment elevation are observed. These diagnoses are difficult to distinguish when phenotypical features are incompletely present. Pericarditis is managed with simple supportive care whereas STEMI requires emergency cardiac catheterization.

Selection and classification of ECGs (cases)

For a prior study,[8] we assembled a pool of 55 ECGs for which 3 attending physicians experienced in ECG interpretation had identified the most likely diagnosis as either pericarditis or STEMI, without being given any clinical information (i.e., their "provisional diagnosis"). We purposely did not include a clinical history to control the experimental setup as tightly as possible; including a history would have added an additional layer of complexity to the diagnosis for each individual ECG (i.e., confusability of the ECG tracing itself, confusability of the history presented, and any interaction between the 2).[9] The lack of history mimics the clinical situation where an expert interprets stacks of ECGs for a hospital or clinic without necessarily interacting with the patient or their chart. These 55 ECGs were then interpreted as pericarditis or STEMI by 78 trainees. For the present study, we selected 10 cases of each provisional diagnosis with varying degrees of "confusability," defined by the degree to which the trainees had preferred one diagnosis over the other. For example, an ECG in which 50% of trainees chose pericarditis and 50% STEMI was considered maximally confusable in the pilot set.

For each of the 20 selected ECGs, the associated patient chart was reviewed to obtain the final discharge diagnosis. The basis of the discharge diagnosis was also reviewed, including the clinical history and physical examination and any investigations including cardiac biomarkers, 2D echocardiogram, angiography, and current, prior, and subsequent ECGs. This determination was done by a cardiology fellow (C.K.W.T.) whose results were reviewed by 2 study investigators (M.V.P., J.D.L.). In 3 cases, the provisional diagnosis of STEMI proved to be pericarditis in the discharge diagnosis. Thus, the final set of 20 ECGs included 13 confirmed as pericarditis and 7 as STEMI.

For our study, each participant read each ECG twice, as part of successive blocks of 20 ECGs. Thus, each participant read 40 total ECGs (see Supplemental Digital Appendix 1 at http://links.lww.com/ACADMED/B322 for a diagram of the study flow). The ECGs are included in Supplemental Digital Appendix 2 at http://links.lww.com/ACADMED/B322.

Participants and study procedures

The study was carried out between May 2, 2017, and March 6, 2018. Participants with a broad range of expertise, including residents and attending physicians, were recruited by email from the NYU Grossman School of Medicine Departments of Emergency Medicine and Cardiology. We also included 2 outside cardiologists. Residents received a $50 honorarium for participation. The study was approved by the NYU Langone Health Institutional Review Board.

Participants rated each ECG on a 5-point scale with the following anchors: "definitely pericarditis," "probably pericarditis," "either pericarditis or STEMI," "probably STEMI," and "definitely STEMI." An example of the task is shown in Supplemental Digital Appendix 3 at http://links.lww.com/ACADMED/B322. Participants were not made aware that the ECGs repeated. The rating task was administered using Qualtrics (Provo, Utah).

Scoring of ECGs (cases)

Cases were scored in 2 ways: dichotomously (correct/incorrect) and using DA-RM (described below). Dichotomous scoring used the interpreted diagnosis the participant selected irrespective of the qualifiers ("probably" or "definitely") and treated the middle response category ("either pericarditis or STEMI") conservatively as STEMI. Areas under the receiver operating characteristic curve (AUCs) are reported below. We also conducted a sensitivity analysis in which the middle ("either") response category was counted as missing. Results were essentially the same (results not shown).

Data analysis

We report Classical Test Theory metrics, including item difficulty and point biserial correlation for each ECG, and test–retest reliability.[10] We anticipated that the reliability value would be modest given the intentional inclusion of confusable cases and the low number of cases overall.

In our analysis, we focused on ECG case characteristics and clinician diagnostic characteristics. To describe the ECG cases in terms of their typicality of either of the 2 diagnoses, we used the proposed DA-RM (see Supplemental Digital Appendix 4 at http://links.lww.com/ACADMED/B322). The mathematics are the same as those of traditional IRM, in this case a graded response model.[11] We report the ECG case/item theta parameter (also termed "location") in logits (i.e., the natural logarithm of the odds of diagnosing the case according to the logic of the linear scale). A high positive logit score for a case indicates a high probability of a participant diagnosing STEMI on that ECG; a high negative logit score indicates a high probability of a participant diagnosing pericarditis on

that ECG. Items at the center of the scale are maximally confusable. For example, an ECG case with a logit value of +1.0 corresponds to an odds (exp(1.0)) of 2.72 favoring a STEMI diagnosis, indicating an implied probability of being diagnosed as STEMI of 73% (odds/[odds + 1] = 2.72/3.72 = 0.73).

Thus, under this model, each individual rater is considered an exchangeable assessor of the degree to which the given ECG reflected a prototypical STEMI case and each ECG as a probe to be located along a continuum of definitive presentation as STEMI according to the response pattern across all participants. To describe each participant's diagnostic characteristics, the graded response model generated tracelines for each participant (see Figure 2), which are model predictions showing how that participant used the 5 diagnostic categories from "definitely pericarditis" to "definitely STEMI" when rating each ECG along the continuum of confusability between the 2 diagnoses.[3–5] Conceptually, the point at which the traceline for "definitely pericarditis" intersected with the traceline for "definitely STEMI" identified an ECG along the confusability continuum for which that participant would demonstrate equipoise in choosing either diagnosis.[12] Whereas the typical IRM model allows us to determine how responses to an item separate participants by their level of ability, the DA-RM interpretation allows us to evaluate how each participant is able to separate ECGs by their degree of resemblance to STEMI.

Results

Study population and general measures

In all, 37 participants completed the full study procedure: 26 emergency medicine residents (17 junior, 9 senior), 6 attending emergency medicine physicians, and 5 cardiologists. Ten emergency medicine residents responded to at least one case but did not complete the study procedure; their data are not included.

The abilities of the different groups to identify which ECGs had a discharge diagnosis of STEMI followed the expected pattern, with cardiologists (AUC = 0.87; 95% confidence interval: 0.82, 0.92) being more accurate than emergency medicine physicians (AUC = 0.80; 95% confidence interval: 0.78, 0.83; $P < .02$) (see Supplemental Digital Appendix 5 at http://links.lww.com/ACADMED/B322). For emergency medicine physicians, accuracy did not differ by training level (i.e., between junior residents, senior residents, and attending physicians). The most capable individual (a cardiologist) showed a sensitivity of 100%, specificity of 61.4%, and an AUC of 0.96.

The test–retest reliability for the 20-item scale was 0.66. The point biserial correlations of each item were positive, ranging from 0 to 0.4, except for 2 items that had negative values, including Case 18: IR described below, suggesting that most items were consistent with the underlying construct.[10] Item proportion correct scores ranged from 0.08 to 1.0, with 2 items rated correctly (in dichotomous scoring) by all participants over both blocks. However, even for those 2 items, participants' ratings on the 5-point scale differed, making the items useful for IRM.

Twelve participants (all residents) restricted their responses to only 4 of the categories, with 11 never selecting the "definitely pericarditis" category.

DA-RM analyses

An advantage of IRM in general is that it can place both the participants' ratings and the cases on the same scale (see Figure 2). In the following sections, we first describe how the cases fell along the latent scale and then we demonstrate how the same scale can be used to describe participants' characteristics.

ECG cases. The DA-RM scaling of the cases revealed that all 20 cases did indeed provide an acceptable discriminative range across the latent scale, with ECG case overall thresholds ranging from −4.23 logits (98% probability of pericarditis diagnosis) to +4.67 logits (99% probability of STEMI diagnosis) (see Supplemental Digital Appendix 4 at http://links.lww.com/ACADMED/B322). Importantly, 6 of the 7 discharge-diagnosis STEMI cases had the highest case locations; conversely, 6 of the 7 cases with the lowest item locations all proved to be pericarditis.

Case 18: IR was an exception. Although the low logit score (−3.96) suggested a modeled diagnosis of pericarditis, the discharge diagnosis was apical STEMI. Here, an 83-year-old woman presented with major gastrointestinal bleeding as well as chest pain. Her troponin was elevated, and an echocardiogram suggested apical myocardial infarction. The ECG, shown in Supplemental Digital Appendix 2 at http://links.lww.com/ACADMED/B322, demonstrates diffuse ST elevations in most leads, a q-wave indicative of prior myocardial infarction, and borderline PR elevation. Four of the 5 cardiologists recognized the uncertainty in this ECG and classified it as "either pericarditis or STEMI," while the fifth diagnosed it as "probably pericarditis."

Clinician diagnostic characteristics. Using DA-RM, we were able to assess each participant on the same scale as the one determined for the ECG cases. The thresholds for the individual participants who used all 5 categories (from "definitely pericarditis" to "definitely STEMI") are shown in Figure 3, demonstrating where on the continuum each participant placed their thresholds between the 5 categories. No threshold locations discernibly differed between cardiologists and the rest of the participants. However, the variance between the threshold values was smaller for cardiologists compared with others, suggesting higher consistency. For example, between "probably pericarditis" and "either pericarditis or STEMI," cardiologists placed their threshold at −0.51 logits, similar to the rest of participants (−0.39 logits; 95% confidence interval difference −0.25, +0.48), but the cardiologists' standard deviation for the threshold was 0.12 compared with 0.35 for the other participants.

Each participant's tendencies across the scale can be represented using probability tracelines, as shown in Figure 2. A comparison of the tracelines between participants demonstrates variation in decision making. In Figure 4, we compare 2 participants and show how they would be predicted to respond differently to ECGs across different parts of the latent scale.

Discussion

In this proof-of-concept study, we used a version of IRM more closely aligned with how a diagnostic decision is made in clinical practice to calibrate a set of ECGs along a potential continuum of

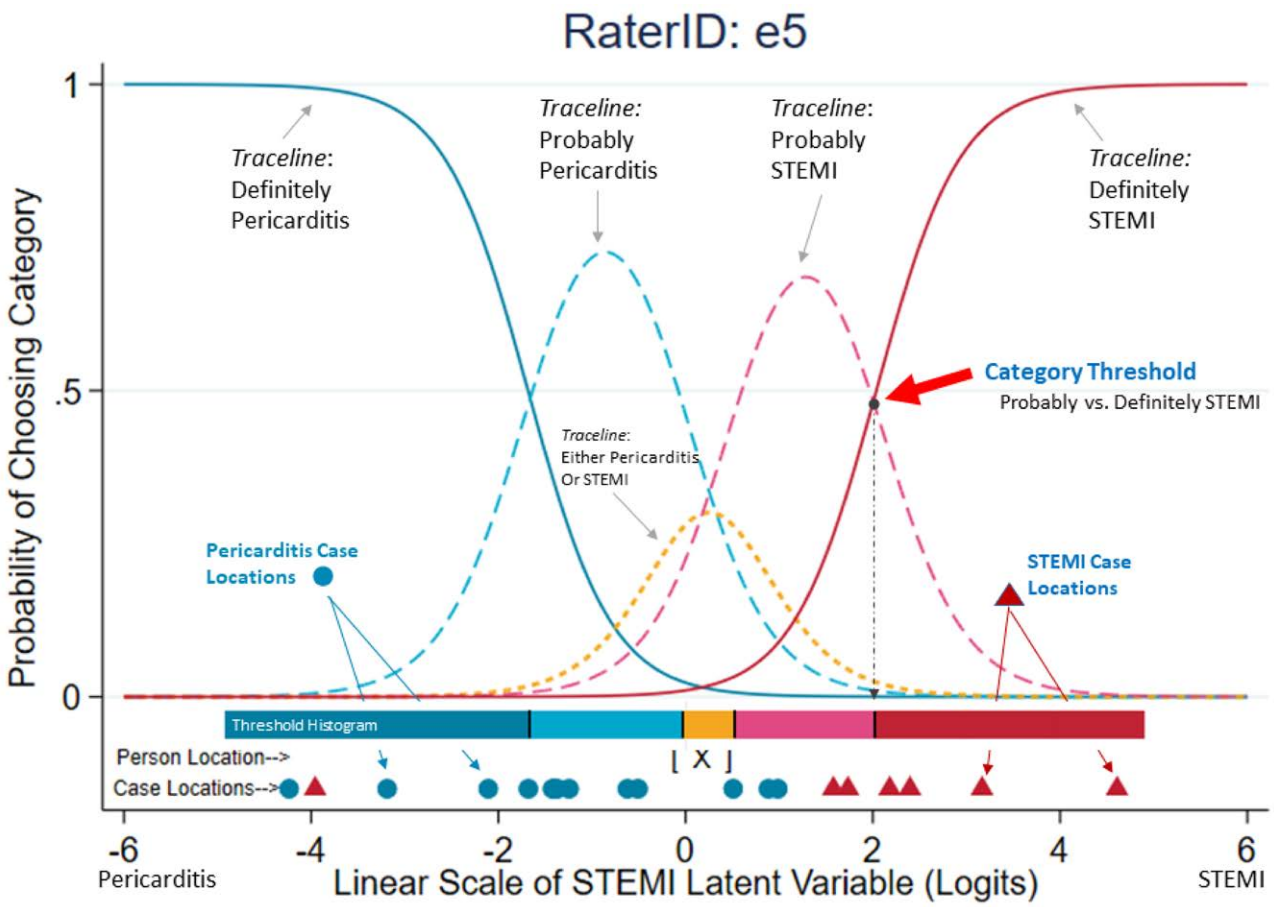

Figure 2 Tracelines for one participant completing 20 ECG cases twice (40 total) in a decision-aligned response model study. Here, we describe each figure element from top to bottom. Subsequent figures use these same definitions. The 5 *tracelines* represent modeled predictions of which category this participant would prefer when confronted with a case at that point on the logit (x-axis) scale. The *category threshold location* is the point on the logit (x-axis) scale where the tracelines for 2 adjacent categories cross, meaning that this participant would be predicted to be equally likely to endorse the categories above or below that threshold for an ECG case at that exact point on the scale. The 5-point scale shown includes 4 category thresholds; we labeled only the "probably STEMI" vs "definitely STEMI" threshold. The *person location* is the tendency or bias of this participant to diagnose cases toward one end of the scale compared with the other; similar to a sensitivity/specificity tradeoff, it is mathematically defined as the point where the top and bottom categories (tracelines) intersect. A person location of zero (as shown here delineated by the "X") indicates a lack of bias in either direction. The *ECG case locations* (blue circles = pericarditis, red triangles = STEMI) are the markers that show the estimated degree to which each of the 20 cases resembles pericarditis (left side) or STEMI (right side), as derived from the responses of all participants. A case at 0 logits would be maximally confusable according to the latent construct, predicted to have equal resemblance to pericarditis and STEMI. The *threshold bar-histogram* is the horizontal 5 color bar that shows which category this participant is most likely to choose for a case at that location on the logit scale. Adjoining changes in color correspond to this participant's category thresholds. In Figure 3, these individual-level bar-histograms are compared for many of the participants in the study. The *x-axis (logit scale)* is the linear psychometric scale whose units correspond to the natural log of the odds of declaring a case STEMI. Positive numbers indicate a higher probability of diagnosing STEMI on that ECG; negative numbers indicate a higher probability of diagnosing pericarditis. Item response modeling generates a participant's category tracelines by conditioning their particular responses with those of all other participants, according to the theoretical response distribution (see text). In the example above, an ECG along the confusability continuum whose logit value is 2.0 would be equally likely to be classified by this participant as "definitely STEMI" vs "probably STEMI" (or lower category). It is possible to calculate the 95% confidence interval (not shown in the figure) for that threshold (1.4, 2.7) indicating the precision of the estimate. Abbreviations: ECG, electrocardiogram; STEMI, ST-elevation myocardial infarction.

confusability between 2 diagnoses and to demonstrate the diagnostic characteristics of each clinician diagnosing the ECGs. Practically speaking, this approach yielded 2 useful insights: (1) We can statistically model the "uncertainty" in a case (i.e., the confusability of one diagnosis with another), and (2) we can better understand a clinician's diagnostic characteristics. By modeling uncertainty, in both the clinician and the case, we can exploit this construct for educational purposes.

While the specific outcome of this study is quite narrow, this approach to assessing diagnosis in a given domain holds great promise for medical educators. In clinical practice, ambiguous or uncertain cases can be common, depending on the clinical context, with our pericarditis-STEMI distinction being only one example of a far broader set of diagnostic phenomena where the underlying construct is a continuum and not a correct/incorrect dichotomy. In traditional assessment approaches, these ambiguous cases would be considered poor test items. Yet, they may be the most salient for clinical practice, where learning to manage uncertainty is part of the development of expertise. DA-RM allows us to identify uncertain cases

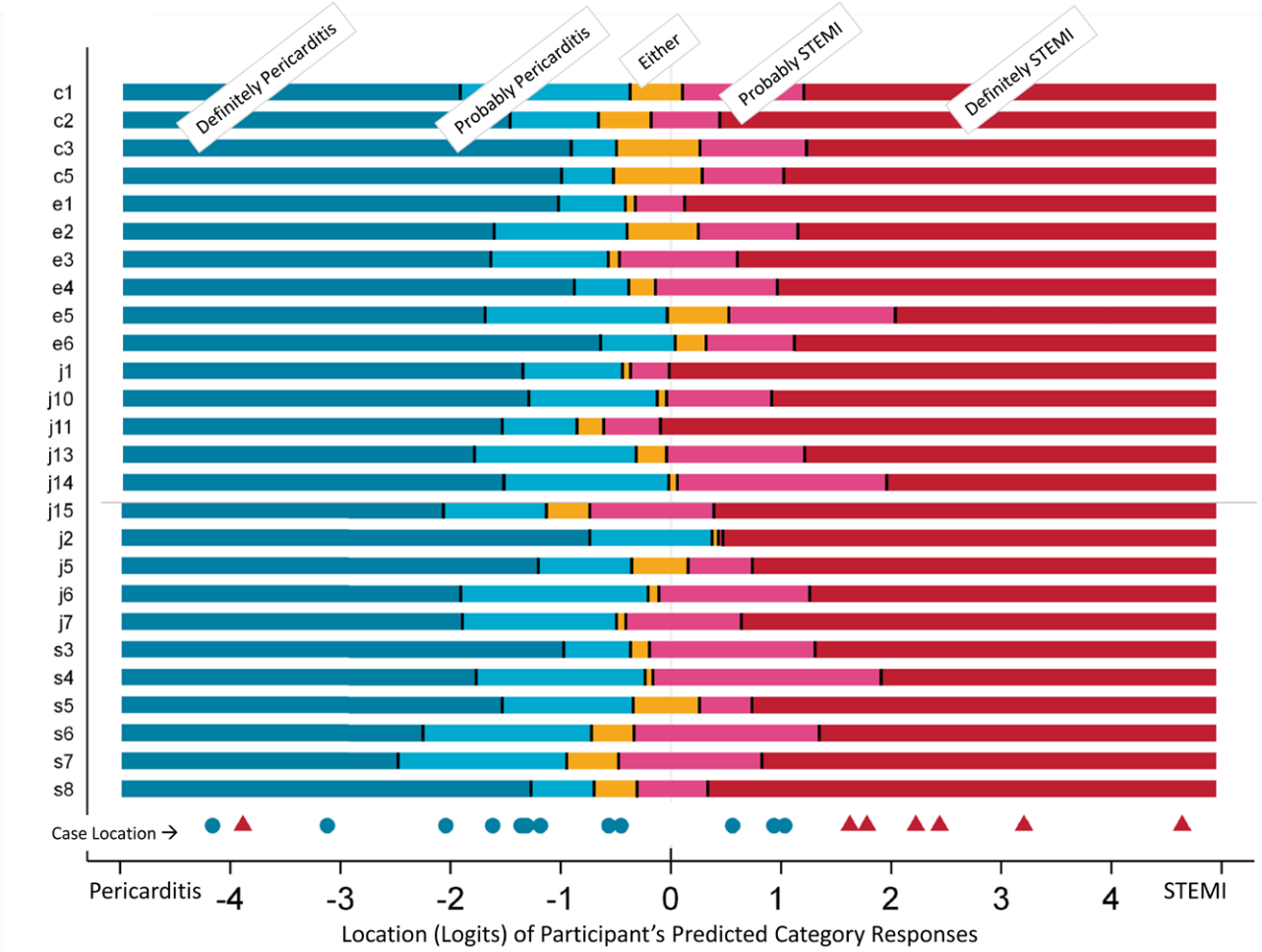

Figure 3 Response thresholds for participants in a decision-aligned response model study. The x-axis scale represents the degree (in logits) to which each participant would be likely to assign the determination of either STEMI (right side) or pericarditis (left side). Each horizontal bar represents a participant, limited to those who used all 5 categories in their responses ("definitely pericarditis" to "definitely STEMI"). The colors show the predicted category that each participant would most likely select for a case at that point on the scale. The colored markers just above the x-axis are the ECG cases placed on the same logit scale, with the color indicating the discharge diagnosis (blue circles = pericarditis, red triangles = STEMI). The STEMI case at the left end (–3.96 logits) of the scale (Case 18: IR) is an exception and is discussed in the text. Certain patterns are apparent from this graph. Use of the categories varies between participants, such that some individuals build in a larger safety margin (assigning borderline pericarditis cases to the STEMI categories [e.g., j15]), and some use the "definitely" category more liberally than others (e.g., j11). The degree to which the middle (yellow) category ("either pericarditis or STEMI") lines up with the zero line indicates the calibration of the participant with respect to a case that is modeled to have a 50% likelihood of either diagnosis. Abbreviations: ECG, electrocardiogram; STEMI, ST-elevation myocardial infarction.

and use them explicitly to understand a clinician's approach to uncertainty.[2,13] Script concordance testing attempts to incorporate this type of acceptable variability into a scoring model[14] but has acknowledged shortcomings.[15]

Of further relevance to medical educators, DA-RM requires that items be chosen for their representativeness of the diagnosis (and perhaps the consequent clinical decision), rather than the traditional approach of selecting items to span ability levels. This approach holds the potential to develop test materials that more closely mimic the range of cases and decisions that a clinician faces in clinical practice, a notable failing of existing ECG assessments.[16] The overall scale demonstrated useful properties, including being able to array the ECG cases on the same linear continuum that demonstrates clinicians' diagnostic characteristics.

Consider how assessment of ECG competence might be enhanced using DA-RM. Based on a test set of ECGs with a calibrated continuum of confusability, novice learners might be tested using only cases modeled to be "definitely pericarditis" or "definitely STEMI," cementing their mental prototypes of the canonical features of both diagnoses.[17] Then, more advanced learners could be presented contrasting cases in the "probably" categories. They would have to consider the following questions: Why is this case only probably pericarditis? What ECG features leave me in doubt about this diagnosis? How would I operationalize that doubt? For learners about to enter practice, the test set could closely mirror the full confusability continuum of the ECGs that a clinician would encounter in clinical practice, with particular focus on the category thresholds that define important decisions, like whether to activate the STEMI alert response.

Furthermore, if the continuum is well modeled, educators would have a granular, quantified understanding

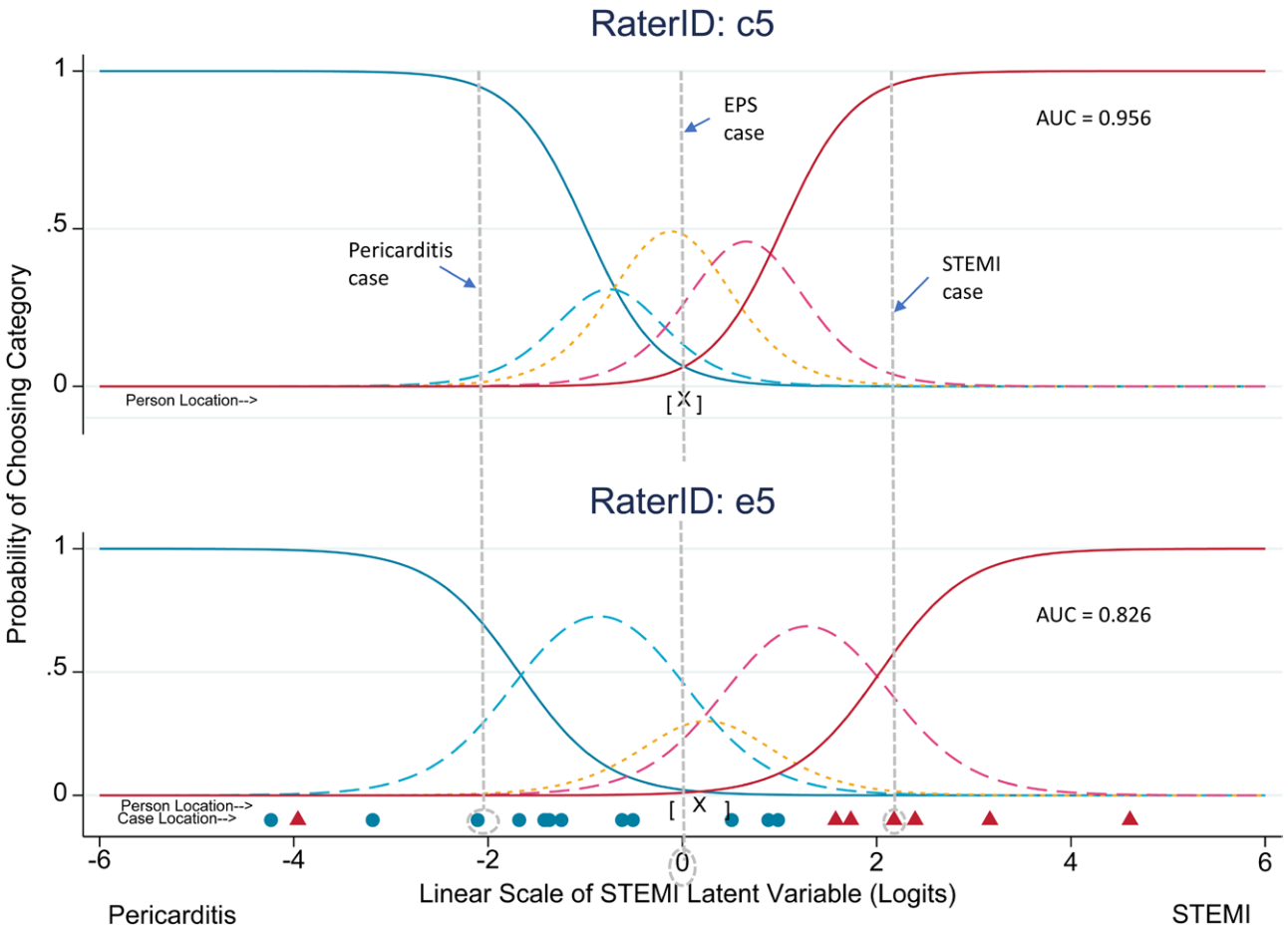

Figure 4 Comparison of decision-aligned response model tracelines for 2 participants. See the Figure 2 legend for an explanation of the figure elements. Rater c5 (upper panel) is much more certain than rater e5 (lower panel) of their diagnoses for the labeled pericarditis and STEMI cases (vertical lines). More specifically, for the pericarditis case at –2.0 logits, rater c5 would almost always use the "definitely pericarditis" category, whereas rater e5 would be predicted to use the "probably" qualifier approximately a third of the time. A similar pattern is seen for the STEMI case at +2.3 logits. The raters also differ in their consideration of a maximally confusable case at 0 logits ("EPS case"), where rater c5 would preferentially choose "either pericarditis or STEMI" in the majority of cases (50% contrasted with 25% for higher and lower categories). For the same case, rater e5 would be predicted to use any of the 3 adjacent categories with equal frequency. Abbreviations: STEMI, ST-elevation myocardial infarction; AUC, area under the receiver operating characteristic curve.

of the decision thresholds for each learner (i.e., which cases represented diagnostic equipoise, which cases were confusing, and which cases could be clearly diagnosed by each learner). Separate from the implications for testing, there are also implications for learning, especially if learners work through a test set in an assessment-for-learning framework. Asking learners to classify cases using a Likert-type rating scale that includes uncertainty would allow educators to accurately identify less certain cases. Indeed, we included a central anchor, "either pericarditis or STEMI," to capture exactly the type of ECG case where it is not possible to tell with certainty which diagnosis is correct based on the information in the ECG alone. By working through ECGs with additional, more detailed feedback as to where the case falls on the continuum (e.g., Figure 2), educators could expect a different type of learning—instead of the learner only considering individual cases in discrete categories, they could develop a mental model of the full continuum.

Even the single ECG case that did not follow the rule, being classified as running counter to the linear scale, could be viewed as an exception that proves the rule. Case 18: IR proved to be a STEMI diagnosis due to a major gastrointestinal bleed; it depicted a (rare) nonfocal type of STEMI where all coronary vessels are hypoperfused, resulting in generalized ST changes resembling pericarditis on the ECG. The fact that the cardiologists in our study were not fooled suggests that the full educational task extends beyond simply teaching learners the scale to also teaching them the underlying logic necessary for dealing with exceptions. An important next step would be to investigate how DA-RM can complement the teaching of approximate rules, thus leading to potentially higher levels of expertise.[18] Investigations of response process, using perhaps "think-aloud" cognitive interviewing methods, would be required to further investigate this idea.[19,20]

Using DA-RM also has implications for clinical practice. The process of placing individuals and items on the same latent scale has a long history in the general psychometric literature.[4,21–23] The application to clinical decisions is more recent. Schwarz used a polytomous IRM to represent the decision as to whether kindergarten students

should be referred for individualized attention.[4] Baldwin and colleagues used a similar psychometric approach in having orthopedic surgeons classify the radiograph appearance of hip fractures.[5] The surgeons used a 4-point ordinal scale of fracture severity to classify the radiographs, demonstrating practice variation in where they placed their decision thresholds. The implication was that the cases would receive a different surgery depending on which surgeon's threshold was used. Zhang and Petersen used similar modeling techniques to compare mammographers on their propensity to identify uncertain cases of breast cancer, suggesting that the technique could have widespread utility in both quality improvement and visual diagnosis training.[3]

In our study, by representing both ECG cases and clinicians' decision thresholds on the same scale, we can statistically predict how a clinician might characterize a given ECG over multiple repetitions. This allows estimation of practice variation across individuals. It also allows a useful target for deliberate practice, with the best clinicians having very precise and reliable thresholds. The perhaps unsettling implication for diagnosis in clinical practice is that, for a given ECG, a competent clinician can be equally likely to choose 1 of 2 or even 3 diagnostic ratings. When a clinician calibrates a high-stakes decision, such as whether to activate a hospital's entire cardiac catheterization response, the need to acknowledge and understand this stochastic (i.e., random) component is important. Furthermore, standard metrics and the "grading" of whether hospitals and clinicians took the "right" action in response to a STEMI diagnosis occur *a posteriori*, without appreciation of the full diagnostic uncertainty involved. This hindsight bias is seemingly at play not only for STEMI recognition but also for other conditions, such as stroke and sepsis. By making it possible to determine clinicians' metrics *a priori*, DA-RM offers the potential of influencing how a given clinician might make such a decision in real time.

Limitations

We undertook this work as a proof-of-concept study to explore DA-RM. Thus, of necessity, we examined a very narrow diagnostic dilemma, with highly controlled test materials (i.e., ECG cases, no history or other information), which limited our ability to directly generalize the results beyond the question we examined. As one peer reviewer pointed out, providing the history for Case:18 IR would have completely changed where that case fell on the modeled diagnostic continuum. While IRM is relatively resistant to participant sampling considerations, the spectrum of clinicians in our study may limit the generalizability of our results, which would need to be verified if applied to other populations, especially if the assessment stakes were higher.[24] The test–retest reliability of our 20-item scale was low, which we took into account in reporting the precision of our point estimates. Among IRM approaches, we selected the graded response model because we wished to capture the interindividual variability in thresholds between Likert-type categories. However, other types of rating scale models might also be suitable, especially the Rasch model, where the interthreshold distances are modeled as being fixed and therefore the same between individuals.[21]

Conclusions

In this study, we demonstrated how a psychometric model allows both quantification of the confusability of a particular ECG and quantification of the skill with which a clinician can distinguish along a continuum of ECGs, representing the full range that can be seen in clinical practice. Our results have broad implications both for skill testing and for learning to manage uncertainty in diagnosis.

Acknowledgments: The authors wish to gratefully acknowledge the contributions of Jackie Gutman, Ilan Reinstein (statistical analysis), Greta Elysee (recruitment), and So Young Oh (survey design). Drs. Jonathan Rubright and Yoon Soo Park gave helpful advice on an early version of the manuscript. The authors also thank the anonymous reviewers whose high level of engagement with the manuscript improved it considerably.

Funding/Support: This work was funded by the U.S. Department of Defense Medical Simulation and Information Sciences Research Program (grant number W81XWH-16-1-0797). The funding source played no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Other disclosures: Martin V. Pusic had full access to all data in this study, takes responsibility for the integrity of the data and the accuracy of the data analysis, and had authority over manuscript preparation and submission. He conducted the data analyses. All authors contributed to obtaining funding, study design, and data collection; reviewed data analyses; revised the manuscript for important intellectual content; and approved the final manuscript.

Ethical approval: This study was reviewed and approved by the Institutional Review Board of NYU Langone Health.

M.V. Pusic is associate professor of emergency medicine, Departments of Pediatrics and Emergency Medicine, Harvard Medical School, Boston, Massachusetts; ORCID: https://orcid.org/0000-0001-5236-6598.

D.A. Cook is professor of medicine and medical education, chair, Mayo Clinic Multidisciplinary Simulation Center Research Committee, and consultant, Division of General Internal Medicine, Mayo Clinic College of Medicine, Rochester, Minnesota; ORCID: https://orcid.org/0000-0003-2383-4633.

J.L. Friedman is assistant professor of clinical medicine, Department of Medicine, Weill Cornell Medical College, New York, New York.

J.D. Lorin is assistant professor, Department of Medicine, NYU Grossman School of Medicine, New York, New York.

B.P. Rosenzweig is associate professor, Department of Medicine, associate director for educational affairs, Leon H. Charney Division of Cardiology, and assistant dean for graduate medical education, NYU Grossman School of Medicine, New York, New York.

C.K.W. Tong is cardiologist and codirector, Heart Failure Services, Surrey Memorial Hospital, Surrey, British Columbia, Canada.

S. Smith is associate professor of emergency medicine, Department of Emergency Medicine, NYU Grossman School of Medicine, New York, New York.

M. Lineberry is associate professor of population health, Department of Population Health, University of Kansas Medical Center and Health System, Kansas City, Kansas; ORCID: https://orcid.org/0000-0002-0177-5305.

R. Hatala is professor, Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada; ORCID: https://orcid.org/0000-0003-0521-2590.

References

1 Tamblyn R, Abrahamowicz M, Dauphinee WD, et al. Association between licensure examination scores and practice in primary care. JAMA. 2002;288:3019–3026.
2 Ilgen JS, Eva KW, de Bruin A, Cook DA, Regehr G. Comfort with uncertainty: Reframing our conceptions of how clinicians navigate complex clinical situations. Adv Health Sci Educ Theory Pract. 2019;24:797–809.
3 Zhang S, Petersen JH. Quantifying rater variation for ordinal data using a rating scale model. Stat Med. 2018;37:2223–2237.

4 Schwarz RD. Trace lines for classification decisions. Appl Meas Edu. 1998;4:311–330.
5 Baldwin P, Bernstein J, Wainer H. Hip psychometrics. Stat Med. 2009;28:2277–2292.
6 Downing S. Item response theory: Applications of modern test theory in medical education. Med Educ. 2003;37:739–745.
7 Billings MS, Deruchie K, Hussie K, et al. NBME Item-Writing Guide. Constructing Written Test Questions for the Health Sciences. Philadelphia, PA: NBME; 2020. https://www.nbme.org/sites/default/files/2020-11/NBME_Item%20Writing%20Guide_2020.pdf. Accessed July 12, 2022.
8 Hatala R, Gutman J, Lineberry M, Triola M, Pusic M. How well is each learner learning? Validity investigation of a learning curve-based assessment approach for ECG interpretation. Adv Health Sci Educ Theory Pract. 2019;24:45–63.
9 Leblanc VR, Norman GR, Brooks LR. Effect of a diagnostic suggestion on diagnostic accuracy and identification of clinical features. Acad Med. 2001;76:S18–S20.
10 Crocker L, Algina J. Item Analysis. In: Introduction to Classical and Modern Test Theory. New York, NY: Holt, Rinehart, and Winston; 1986:311–338.
11 Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika. 1969;34:1–97.
12 Raykov T, Pusic M. Evaluation of polytomous item locations in multicomponent measuring instruments: A note on a latent variable modeling procedure [published online ahead of print March 2022]. Educ Psychol Meas. doi :10.1177%2F00131644211072829.
13 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;35:232–243.
14 Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: From theory to practice: AMEE guide no. 75. Med Teach. 2013;35:184–193.
15 Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47:1175–1183.
16 Cook DA, Oh SY, Pusic MV. Assessments of physicians' electrocardiogram interpretation skill: A systematic review. Acad Med. 2022;97:603–615.
17 Brush JE, Jr, Sherbino J, Norman GR. Diagnostic reasoning in cardiovascular medicine. BMJ. 2022;376:e064389.
18 Yoon JS, Boutis K, Pecaric MR, Fefferman NR, Ericsson KA, Pusic MV. A think-aloud study to inform the design of radiograph interpretation practice. Adv Health Sci Educ Theory Pract. 2020;25:877–903.
19 Ericsson KA. Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In: Ericsson KA, Charness N, Feltovich P, Hoffman RR, eds. Cambridge Handbook of Expertise and Expert Performance. New York, NY: Cambridge University Press; 2006:223–242.
20 Crandall B, Klein G, Hoffman RR. Working Minds: A Practitioner's Guide to Cognitive Task Analysis. Cambridge, MA: MIT Press; 2006.
21 Andrich D, Marais I. Invariance of comparisons—Separation of person and item parameters. In: A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences. Singapore: Springer; 2019:89–96.
22 Wright BD, Masters GN. Rating Scale Analysis. Rasch Measurement. Chicago, IL: Mesa Press; 1982. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1001&context=measurement. Accessed July 12, 2022.
23 Rasch G. On general laws and the meaning of measurement in psychology. Berkeley Symp Math Stat Probab. 1961;4.4;321–333.
24 DeVellis RF, Thorpe CT. Scale Development: Theory and Applications. 5th ed. New York, NY: Sage Publications; 2022.