Advances in Health Sciences Education (2019) 24:45–63
https://doi.org/10.1007/s10459-018-9846-x

How well is each learner learning? Validity investigation of a learning curve-based assessment approach for ECG interpretation

Rose Hatala[1] · Jacqueline Gutman[2] · Matthew Lineberry[3] · Marc Triola[2] · Martin Pusic[2,4]

Received: 18 September 2017 / Accepted: 26 July 2018 / Published online: 31 August 2018

Abstract

Learning curves can support a competency-based approach to assessment for learning. When interpreting repeated assessment data displayed as learning curves, a key assessment question is: "How well is each learner learning?" We outline the validity argument and investigation relevant to this question, for a computer-based repeated assessment of competence in electrocardiogram (ECG) interpretation. We developed an on-line ECG learning program based on 292 anonymized ECGs collected from an electronic patient database. After diagnosing each ECG, participants received feedback including the computer interpretation, cardiologist's annotation, and correct diagnosis. In 2015, participants from a single institution, across a range of ECG skill levels, diagnosed at least 60 ECGs. We planned, collected and evaluated validity evidence under each inference of Kane's validity framework. For Scoring, three cardiologists' kappa for agreement on correct diagnosis was 0.92. There was a range of ECG difficulty across and within each diagnostic category. For Generalization, appropriate sampling was reflected in the inclusion of a typical clinical base rate of 39% normal ECGs. Applying generalizability theory presented unique challenges. Under the Extrapolation inference, group learning curves demonstrated expert–novice differences, performance increased with practice and the incremental phase of the learning curve reflected ongoing, effortful learning. A minority of learners had atypical learning curves. We did not collect Implications evidence. Our results support a preliminary validity argument for a learning curve assessment approach for repeated ECG interpretation with deliberate and mixed practice. This approach holds promise for providing educators and researchers, in collaboration with their learners, with deeper insights into how well each learner is learning.

Keywords ECG interpretation · Longitudinal assessment · Learning curve · Validity evidence

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10459-018-9846-x) contains supplementary material, which is available to authorized users.

✉ Rose Hatala
rhatala@mac.com

Extended author information available on the last page of the article

Introduction

Competency-based education requires a shift in assessment approaches from biopsies of a learner's performance at a single point in time to repeated assessment of performance over time to ensure that competency is attained and maintained (Schuwirth and van der Vleuten 2011a). Increased integration of assessment into the learning environment may facilitate this shift, so that while discrete moments of assessment still occur (traditional assessment of learning), there is an increased emphasis on ongoing assessment for learning (Schuwirth and van der Vleuten 2011b). Examining the shape of learners' learning curves across their repeated practice of some aspect of knowledge and/or skill may be an ideal method of interpreting assessment data for competency-based medical education. Specifically, the use of repeated assessment compared against a competency standard and expected progression rates, all integrated with immediate feedback on performance, models assessment for learning (Pusic et al. 2015a).

Typical learning curves plot performance along a y-axis against effort on an x-axis (e.g., time, number of cases) and have 4 components: (a) a y-intercept indicating baseline performance; (b) a rapid upward initial slope indicating an efficient early learning phase; (c) a slower incremental learning phase where more difficult concepts are mastered; and (d) an upper asymptote of maximal performance (Pusic et al. 2015a). Both individual learner and group-level (typically based on level of expertise) learning curves may be plotted and analyzed either descriptively or mathematically (Pusic et al. 2016).

While surgical domains have a long history of using learning curves during skill acquisition (Ramsay et al. 2001), learning curves have been underexplored for use with cognitive skill development aside from a series of studies examining learning curves during pediatric ankle radiograph interpretation in a computer-based learning application. Those studies established learning curves and competency standards for different levels of learners (Pusic et al. 2011), emphasized the importance of fidelity to normal/abnormal image ratios in real clinical practice (Pusic et al. 2012a), and reported the development of improved self-monitoring during practice (Pusic et al. 2015b).

Approaches such as the pediatric ankle radiograph learning system offer a potential advantage over current approaches to learning to interpret visual information in clinical settings, where learning is often fragmented (i.e., occurring in some clinical settings but not others), limited by the few cases available to an individual learner, and limited by a lack of feedback on performance (Ericsson 2015). Thus, in clinical settings, the range of ability required to interpret visual material is both enormous and arbitrary whereas a computer-based system may provide a more controlled learning environment. Furthermore, a computer-based system capturing both process metrics (e.g. time per page, sequence through the material) and outcome metrics (e.g. diagnosis, confidence) facilitates the integration of learning analytics to understand both learner and learning system performance, thus providing richer information to both learners and educators (Pecaric et al. 2017).

Similar to radiographs, ECG interpretation is another visual perception task that is difficult to learn in the clinical setting and where a computer-based system may be advantageous (Fent et al. 2015). Though numerous approaches to teaching ECG interpretation have been described, they are frequently time-limited, provide few practice examples to learners, and assess learning at discrete moments in time (e.g., at an immediate or delayed post-test) (Fent et al. 2015; Rourke et al. 2018). Furthermore, validity evidence to support the assessment approaches used in the literature is lacking (Rourke et al. 2018). In recent review papers summarizing the ECG instructional literature, no single approach to or format of ECG teaching was superior when compared each to the other (Fent et al. 2015; Rourke et al. 2018), although eliciting learner performance and providing feedback led to larger learning gains as assessed from pre- to post-test (Rourke et al. 2018). A computer-based ECG learning system has been described, where learners practice 15 diagnoses over 75 ECGs. Students' ECG interpretation skills after implementation of this computer-based system were superior to an historical cohort taught with classroom-based methods (Chudgar et al. 2016). However, while this system provided an improved opportunity for practice with feedback, learner performance was only assessed with a 10-item post-test.

Addressing ECG competency is an important clinical issue as ECG diagnostic errors are made at all levels of learners and clinicians (Jablonover et al. 2014; Salerno et al. 2003a). There is no standard, reliable, competency-based assessment addressing ECG diagnostic accuracy (Salerno et al. 2003b). In addition, there are no evidence-based data as to how many ECGs are necessary to achieve competence and expert recommendations vary widely (Salerno et al. 2003a). Computerized interpretation is not a panacea; depending upon the diagnosis, studies suggest computer accuracy is approximately 90% (Guglin and Thatai 2006; Shah and Rubin 2007). The implications for missed abnormalities are serious.

The implementation of an ECG computer-based program based on a high-volume of cases, immediate feedback on performance and examination of learners' learning curves, may aid in predicting a learner's expected trajectory of learning with consistent practice. Individual learning curves may be examined for formative purposes, aiding the learner and educator in understanding the learning that is taking place by examining the shape of the curve. In an ideal world, everyone learns to the asymptote. However, competency thresholds may be established to put boundaries on learning. Capturing group-level learning curves could contribute to competency standard-setting, and subsequently suggest the estimated volume of ECGs required to attain that competence. Standard-setting committees could decide if expert-level performance defined competency [as is common in surgical skills training captured by learning curves (Ramsay et al. 2001)] or if the competency standard should be set at a different threshold.

Any new assessment approach requires validation (Cook et al. 2015). The validity investigation process focuses on examining the key assumptions and inferences that link the assessment tool with its intended use. While various validity frameworks are available, we prefer the approach proposed by Kane (2013) as it encourages educators to describe the educational decision that the assessment addresses. The approach not only outlines the types of evidence that may be relevant to support the validity argument [as does Messick's validity framework (Messick 1989)], but also helps to prioritize among the various evidence sources in order to build a coherent validity argument (Cook et al. 2015; Kane 2013).

The current study outlines the validity investigation process and validity argument, using Kane's framework, for a computer-based learning system with repeated assessment of ECG interpretation. The program was designed using the principles of deliberate practice and immediate feedback with a repeated, learning curve-based approach to conceptualizing performance and drawing inferences about learners. We outline, collect and examine the validity evidence relevant to this repeated assessment approach. In addition to providing the direct evidence supporting learning curves as an assessment of ECG interpretation skills, the current study is the first to our knowledge to articulate a validity argument for a repeated assessment approach using learning curves.

Methods

Development of the computer-based ECG system

Selection and annotation of ECGs for the on-line learning system

The goal of the ECG image bank underlying the computer program is to create an authentic representation of ECGs encountered in clinical practice (Pusic et al. 2012a). Using an iterative Delphi technique which included a panel of expert cardiologists at the University of British Columbia (UBC) as well as the study investigators, we chose a guiding clinical context of plausible ECG diagnoses in patients who present with chest pain. Based on consensus discussion by the panel, the following ECG diagnoses were included in the digitized image bank: normal; ST-elevation myocardial infarction; myocardial ischemia; pericarditis; bundle branch blocks; and ventricular hypertrophy. Arrhythmias were excluded from these diagnoses (i.e. all examples were in normal sinus rhythm) as a narrower differential diagnosis list was felt to be more consistent with the abilities of early undergraduate medical students. These diagnoses were also consistent with the diagnoses used in a previous investigation of ECG instruction (Hatala et al. 2003).

The study was approved by the UBC Behavioural Research and NYU Research Ethics Boards. We subsequently amassed a collection of 299 anonymous ECGs representing each diagnostic category using the MUSE© electronic patient ECG database at a tertiary care hospital, including inpatient, outpatient and emergency department settings. Each ECG was read by three independent cardiologists: (1) the original staff cardiologist within the hospital's electronic database; (2) a second, masked staff cardiologist who interpreted each ECG and who digitally annotated each ECG by circling key features on the ECG that lead to the diagnosis; and (3) a third cardiologist who interpreted each ECG. Disagreements were resolved by consensus and 7 ECGs were removed from the final set as they contained a significant arrhythmia.

Development of the on-line learning system

The refined set of 292 ECGs, including the cardiologist's annotations, were embedded into a Web application developed using the Python Django programming framework and a MySQL database. The set contained 113/292 (39%) normal ECGs. This proportion represents a balance between learning efficiency (which requires a higher proportion of abnormals) versus fidelity to the base prevalence of 60–75% normal in actual clinical care (which requires a higher proportion of normals) (Ashley et al. 2000; De Bacquer et al. 1998; Pusic et al. 2012a). Click-level data were logged for each user's actions at every step of each ECG into a MySQL database. An example ECG is available at: https://education.med.nyu.edu/ecg/example.

Secure entry was ensured via a username and password given to each participant. The software tracked their progress through the ECGs and recorded every response to the database. ECGs were randomly presented in a mixed order which was different for each participant. For each ECG, the participant was presented the ECG without computerized interpretation and was instructed to check rate, rhythm, and axis; note abnormal features; and consider the diagnosis. Proceeding to the second screen, the participant provided a confidence rating on their presumptive diagnosis by rating the ECG on a 4-point scale from

Definitely Abnormal to Definitely Normal. Proceeding to the third screen, the participant selected the final diagnosis for the ECG based on a drop-down list. Entering their diagnosis led them to the final screen, where customized feedback on their responses was provided in the form of the ECG with computerized interpretation, expert annotation, and a description of the correct diagnosis. Participants were given the chance to review this information before moving on to the next ECG; thus, deliberate practice [practicing a specific aspect of performance with repetition and reflection (Ericsson et al. 1993)] and mixed practice (practicing all ECG diagnoses in mixed sequence as opposed to practicing a set of one diagnosis, then a set of the next diagnosis, etc.) were the dominant instructional strategies (Hatala et al. 2003). Participants could access the on-line system at any time and log-in over multiple occasions with their progress book-marked. Participants were not restricted from using other ECG learning materials in addition to the on-line system.

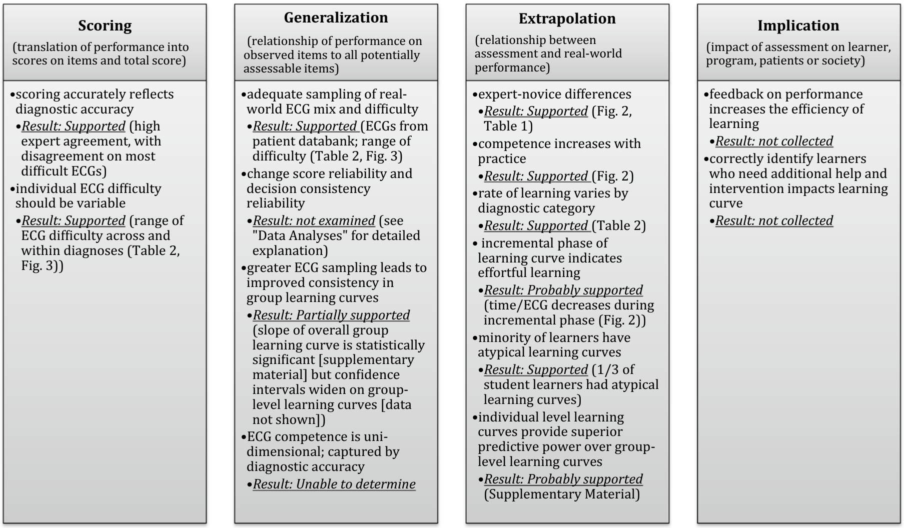
Implementation of the on-line learning system

In 2015, the on-line learning system was implemented across undergraduate and postgraduate medical education programs at UBC. We purposely sampled across a range of ECG 'expertise' from novice (i.e., no prior ECG exposure) to expert (i.e., frequently interpret ECGs in clinical practice). The system was implemented among first-year (pre-clinical) undergraduate students, third- and fourth-year (clinical) undergraduate students during clerkship and UBC PGY-1 internal medicine residents during their summer 'boot-camp' to prepare them for residency training. A voluntary cohort of UBC cardiology fellows were recruited to serve as the expert group. While accessing the on-line system and working through a minimum of 30 ECGs was mandatory for the first year medical students and first year residents, all participants in this report voluntarily consented to allow their data to be used for study purposes.

Validation process

Kane's approach to assessment validity investigation begins with outlining the educational decision which the assessment is intended to support (Cook et al. 2015; Kane 2013). Next the intended use argument (IUA) is delineated, where the key assumptions underlying the assessment scores are outlined a priori. Validity evidence is then gathered under four categories of inference: Scoring, Generalization, Extrapolation and Implications (see Fig. 1 for definitions of each inference). The evidence is examined as to how well it supports the intended use argument and the relevant assessment decisions.

In a model of repeated assessment using learning curves, a key assessment decision is the educator's answer to the question: "How well is each learner learning?" Within that broad question, several sub-questions might be asked. Examining the slope of the learning curve can help answer "Is the rate/effort/efficiency of learning appropriate?" Examining the incremental phase, we might further ask "Has the learner sustainably achieved the intended level of competence?", "Is this learner's curve so flat that altering the learning environment is indicated?", and "Comparing achievement between learners, is there sufficient consistency among this group of learners with respect to competence?" Answering each of these questions could lead to an action on the part of the educator. Within these questions, the focus of the educator and learner is on the shape of the curve and what that implies about the learning trajectory towards a competency standard, rather than solely focussing on attainment of the competency. The richness of



Scoring

(translation of performance into scores on items and total score)

- scoring accurately reflects diagnostic accuracy
  - *Result: Supported* (high expert agreement, with disagreement on most difficult ECGs)
- individual ECG difficulty should be variable
  - *Result: Supported* (range of ECG difficulty across and within diagnoses (Table 2, Fig. 3))

Generalization

(relationship of performance on observed items to all potentially assessable items)

- adequate sampling of real-world ECG mix and difficulty
  - *Result: Supported* (ECGs from patient databank; range of difficulty (Table 2, Fig. 3)
- change score reliability and decision consistency reliability
  - *Result: not examined* (see "Data Analyses" for detailed explanation)
- greater ECG sampling leads to improved consistency in group learning curves
  - *Result: Partially supported* (slope of overall group learning curve is statistically significant [supplementary material] but confidence intervals widen on group-level learning curves [data not shown])
- ECG competence is uni-dimensional; captured by diagnostic accuracy
  - *Result: Unable to determine*

Extrapolation

(relationship between assessment and real-world performance)

- expert-novice differences
  - *Result: Supported* (Fig. 2, Table 1)
- competence increases with practice
  - *Result: Supported* (Fig. 2)
- rate of learning varies by diagnostic category
  - *Result: Supported* (Table 2)
- incremental phase of learning curve indicates effortful learning
  - *Result: Probably supported* (time/ECG decreases during incremental phase (Fig. 2))
- minority of learners have atypical learning curves
  - *Result: Supported* (1/3 of student learners had atypical learning curves)
- individual level learning curves provide superior predictive power over group-level learning curves
  - *Result: Probably supported* (Supplementary Material)

Implication

(impact of assessment on learner, program, patients or society)

- feedback on performance increases the efficiency of learning
  - *Result: not collected*
- correctly identify learners who need additional help and intervention impacts learning curve
  - *Result: not collected*

Fig. 1 Assumptions underlying each inference of Kane's validity framework for the ECG learning curve-based assessment approach. The intended use argument is "How well is a learner learning?" Each column reflects one inference in Kane's validity framework. Under each inference we present the specific assumptions outlined prior to data collection, and whether or not the assumptions were ultimately supported by the data. For a more conceptual description and discussion of each inference see (Kane 2013; Cook et al. 2015)

this IUA is in contrast to a single biopsy of performance where the educator is only able to address the question "*Did* each learner learn?".

However, before we can answer these questions, we must determine if the validity argument holds up to scrutiny. Kane's framework helps us to plan the research required to support our repeated assessment of ECG learning. Our validity investigation is outlined here in brief, and in more detail in Fig. 1. We began by specifying the IUA as above: "How well is each learner learning?". We subsequently outlined the assumptions under each inference which could support this IUA, beginning with Scoring and with diagnostic accuracy as our primary scoring measure. If our approach to Scoring was accurate, we would expect experts to agree on the diagnosis for each ECG and we would expect a range of ECG difficulties. Under Generalization, we would expect adequate sampling of ECGs and high reliability of the scores in order to support decisions as to how well a learner was learning under the specific study conditions. Under Extrapolation, we would expect relationships between learning curve and real-world performance, such as expert-novice differences for the group learning curves and typicality of the individual learning curves [i.e. typicality, as defined by the 4 key phases of y-intercept, efficient learning phase, incremental learning phase and upper asymptote (Pusic et al. 2015a)]. Finally, examination of the impact of the repeated assessment on the learner, the educational system or patients would form the Implications part of the validity argument. However, as the assessment was being used for research purposes, we anticipated being unable to collect meaningful Implications evidence.

The validity evidence required to support each of these inferences form our research hypotheses and lay out the plan for data analysis. As is apparent from the breadth of the assumptions in Fig. 1, it is not feasible to gather all the possible validity evidence within a single study. Thus, our plan for data collection and analysis was guided by Fig. 1.

Data collection

We collected demographic data for each participant regarding level of training and prior ECG experience. The on-line system captured time on task, number of times the system was accessed, number of ECGs completed and diagnostic accuracy per ECG. Our primary outcome measure was diagnostic accuracy; a secondary outcome measure was time on task.

Data analyses

All analyses are based on participants who completed a minimum of 60 ECGs, which we determined was the minimum number of ECGs required to establish a stable learning curve (unpublished data). We used descriptive analyses for the demographic data and we calculated the median number of ECGs interpreted and the time per ECG by level of learner. For each learner, we determined their sensitivity and specificity in classifying ECGs as normal or abnormal, and computed the average sensitivity and specificity (along with standard errors) among medical students, first-year residents, and cardiology fellows (Genders et al. 2012).

We modelled the group learning curve using a mixed-level hierarchical logistic model where the dependent variable was the log odds of correctly diagnosing the ECG while the independent variables were the sequence number (1–60) of the ECG item (log-transformed) and the ECG item difficulty. The details of the model are presented in the Supplementary Material.

To gather validity evidence under the Scoring inference, we calculated the Cohen's kappa on the three experts' agreement on the diagnosis for each ECG based on the six included diagnostic categories plus 'other diagnosis'. A Welch's *t* test was used to compare item difficulty on ECGs with and without full expert consensus. We examined the correlation between scoring partially correct versus fully correct diagnoses (i.e. awarding partial marks if the participant gave a diagnosis of 'ischemia, anterior leads' when the correct diagnosis was 'ischemia, inferior leads').

We computed the item difficulties using the percent correct for each ECG.

The Generalization inference presents unique challenges such that we are unable in the current data collection to provide strong empirical evidence for or against the claims within it. Generalization would be supported if scores under our particular set of testing conditions corresponded well to those that might be collected under different testing conditions. Each learner encountered a limited, random sample of ECGs (i.e., particular items), which correspond to an even more limited set of underlying diagnoses. Any noise associated with the random sampling along these or similar facets would threaten our intended interpretations. For instance, an educator might infer from a flat learning curve that the learner is no longer learning and intervention is required. If the flat slope was simply random sampling noise, the inference would be invalid. This can be conceptualized as an issue of *change score reliability,* i.e., the reliability with which one estimates the difference between performance across two points in time (Brennan 2001). More specifically, the learning curve-based approach calls for a marrying of change score reliability with conceptions of *decision consistency reliability* (Livingston and Lewis 1995; Webb et al. 2006) because it is not essential that the exact value of the change score (or slope) be known, but rather that the *decision* in response to that difference be reliable.

One way of approaching these reliability concepts is with advanced generalizability theory, and further psychometric theoretical work to combine them is needed, as well as a data collection appropriately designed to systematically represent the relevant facets. However, we can impute from regression modelling that if the slope of the group learning curve (i.e., the improvement in performance with each successive ECG) for the overall group of learners is statistically significant with a relatively precise confidence interval, then the majority of the learners in our set are reliably learning. That is, the items chosen from the pool of all possible items are indeed the ones that allow this set of learners to learn. This is analogous to the testing context where the overall internal consistency of a test (e.g. Cronbach's alpha) is Generalization evidence that the test is functioning as it should (Kane 2013).

An additional challenge with the Generalization inference is related to the Scoring assumption that diagnostic accuracy accurately captures ECG competency. This treats ECG competency as being uni-dimensional, while it is possible that it is multidimensional and thus any single y-axis measure across all ECGs is insufficient. For example, perhaps there are a series of important concepts (e.g., recognizing normal ECGs, understanding axis, etc.) that learners can only attain after reading a certain volume of specific ECGs. If so, then diagnostic accuracy plotted on the y-axis as a continuous variable would not fully reflect the development of ECG interpretation skills at an individual level. For similar reasons to those that prevented us from computing generalizability theory analyses, the data were also not structured to be suitable for factor analysis.

For the Extrapolation inference, we generated the learning curves for a) accuracy and b) time as raw moving averages across the learners' last 20 ECGs, both raw and adjusted for the difficulty of the ECG item. We determined the significance of an overall group-level learning curve model as described above. Knowing we did not have enough observations to regression model individual-level learning curves, two experts inspected the descriptive individual-level (moving average) curves for each medical student and rated the curves as 'typical' or 'atypical' by visual inspection (Pusic et al. 2011). The assumption that each learner would have a typical learning curve is based on a number of theoretical foundations including: (a) test-enhanced learning (active questioning fosters learning) (Larsen et al. 2008); (b) feedback-enhanced learning (feedback fosters learning) (Shute 2008); and (c) learning curve theory (in an effective learning environment, individuals learn asymptotically) (Pusic et al. 2015a).

We also compared a series of nested hierarchical logistic regression models (see the Supplementary Material for details), beginning with the null hypothesis and successively adding first a parameter for random variability of the y-intercept and subsequently allowing both y-intercept and slope to vary randomly between individuals.

We did not collect Implications evidence.

Results

Seventy-eight participants (out of 444 learners who logged onto the system) diagnosed a minimum of 60 ECGs (Table 1). The group learning curves by level of expertise are presented in Fig. 2 and Table 1.

Table 1 ECG group learning curve data and analyses

Expertise	n	Median number of ECGs read [25–75%]	Median time per ECG [25–75%]	Average sensitivity [95% CI]	Average specificity [95% CI]	Average y-intercept (accuracy)
Medical students	63	93 [73.5–134.5]	41.5 s [32.5–66.5 s]	90.6% [89.2–92.0%]	80.3% [77.8–82.7%]	48.5% (SD 12.5%)
PGY-1 residents	10	85.5 [73.5–93]	68.2 s [32. 63–92 s]	90.2% [88.0–92.4%]	84.0% [76.7–91.3%]	53.6% (SD 11.3%)
Cardiology fellows	5	214 [189–283]	14 s [12–15 s]	93.9% [91.7–96.0%]	94.4% [92.1–96.6%]	68.9% (SD 14.5%)

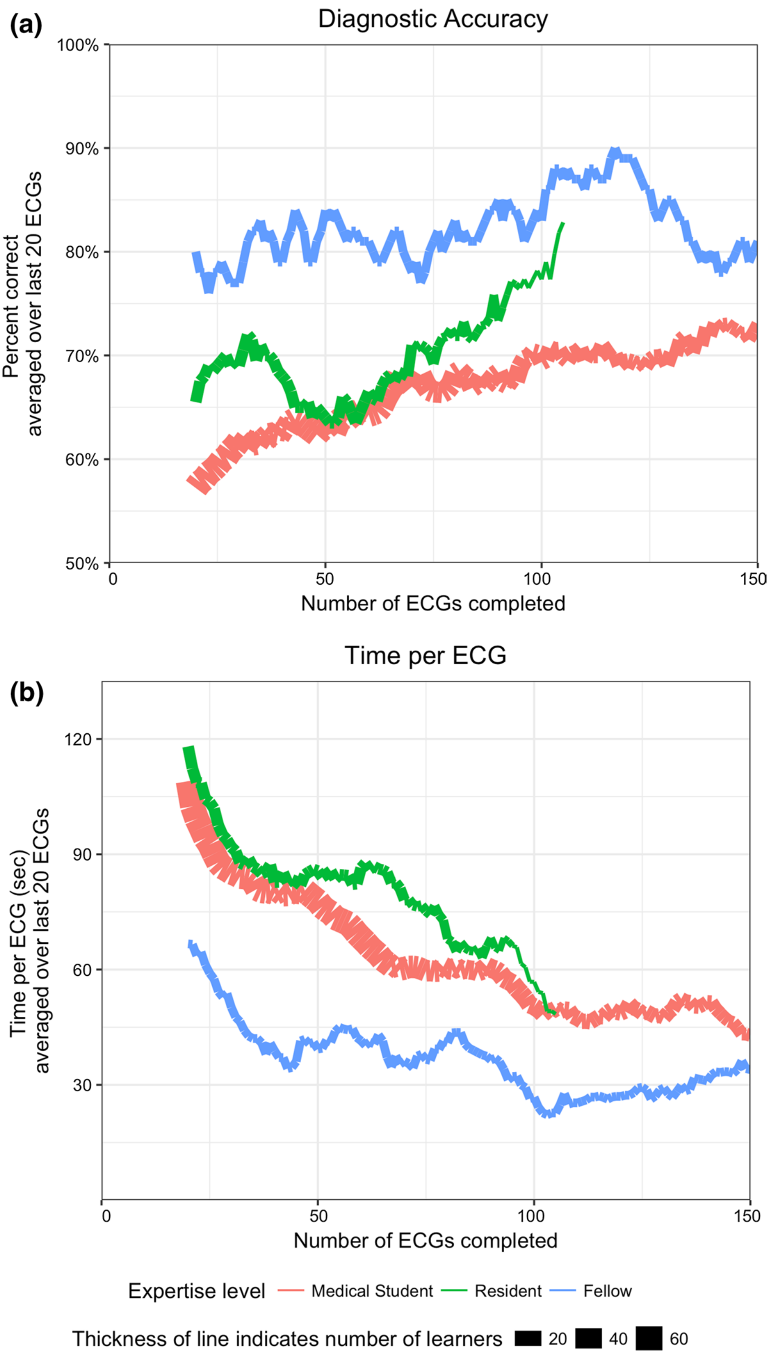


Fig. 2 Group learning curves. The two panels depict the group learning curves for each level of expertise (blue = fellow, green = resident, red = student) using two outcome measures: **a** diagnostic accuracy; **b** time per ECG. (Color figure online)

Scoring inference

The correlation between diagnostic accuracy measured in terms of partial score versus full score was 0.94, so partial scores are reported throughout. If Scoring is accurate, and each ECG has a correct diagnosis, then agreement between experts on the correct diagnosis should be high. The Cohen's kappa for the 3 raters across the 292 ECGs was 0.92. Furthermore, there was a significant difference in difficulty for the ECGS by the level of agreement between the three expert raters, F(2, 290) = 9.82, $p < .001$. A post hoc contrast showed the difficulty ratings of 266 ECGs for which all three experts agreed on the diagnosis was significantly less than the 26 ECGs for which at least one of the experts disagreed, with an estimated difference in difficulty of − 1.48 logits (95% CI − 2.43 to − .54). This corresponds to an average increase in difficulty from approximately the 49th percentile of ECG difficulty for an ECG with full rater agreement to the 87th percentile of ECG difficulty for an ECG on which at least one of the expert raters disagreed.

Further support for the Scoring inference is demonstrated in Fig. 3, where it is apparent that there is a range of ECG difficulty across and within each diagnostic category.

Generalization inference

Support for the Generalization inference rests on sampling and reliability. As outlined in the Methods, we were unable to run Generalizability analyses. At a descriptive level, we would expect that greater sampling of ECGs leads to more stable individual and group learning curves. The individual learning curves demonstrated in Fig. 4 could be visually interpreted as demonstrating large variability, but generally flattening out into a coherent and typical learning curve pattern as a learner reads more ECGs. Thus the reliability of an individual's current state in the learning curve may be improving across ECGs. Furthermore, as the number of ECGs completed increases, the consistency among learners within a level of expertise would be expected to increase and could be reflected in narrowing of the confidence intervals around the group learning curve as participants read more ECGs. However, this is did not occur, as the confidence intervals for the group-level learning curves do not narrow as learners complete more ECGs, likely due to confounding by the decreasing number of learners who completed a greater number of ECGs (data not shown).

Extrapolation inference

The Extrapolation Inference would be supported by demonstrating that group learning curves by level of expertise are distinct and hence reflect varying levels of expertise. As demonstrated in Fig. 2a, different levels of expertise are reflected in the group learning curves. The group learning curves demonstrate the expected pattern of a rapid upward slope indicating efficient learning (with a statistically significant slope for the entire population in the regression model (see Supplementary Material)), an incremental phase indicating a slower rate of learning and an upper asymptote of maximal performance. Table 2 demonstrates that the rate of learning varies by ECG diagnostic category (because we expect that some diagnoses are inherently harder than others).

Further support for the Extrapolation inference would be reflected in the incremental phase of the diagnostic accuracy learning curve, which should indicate ongoing effortful

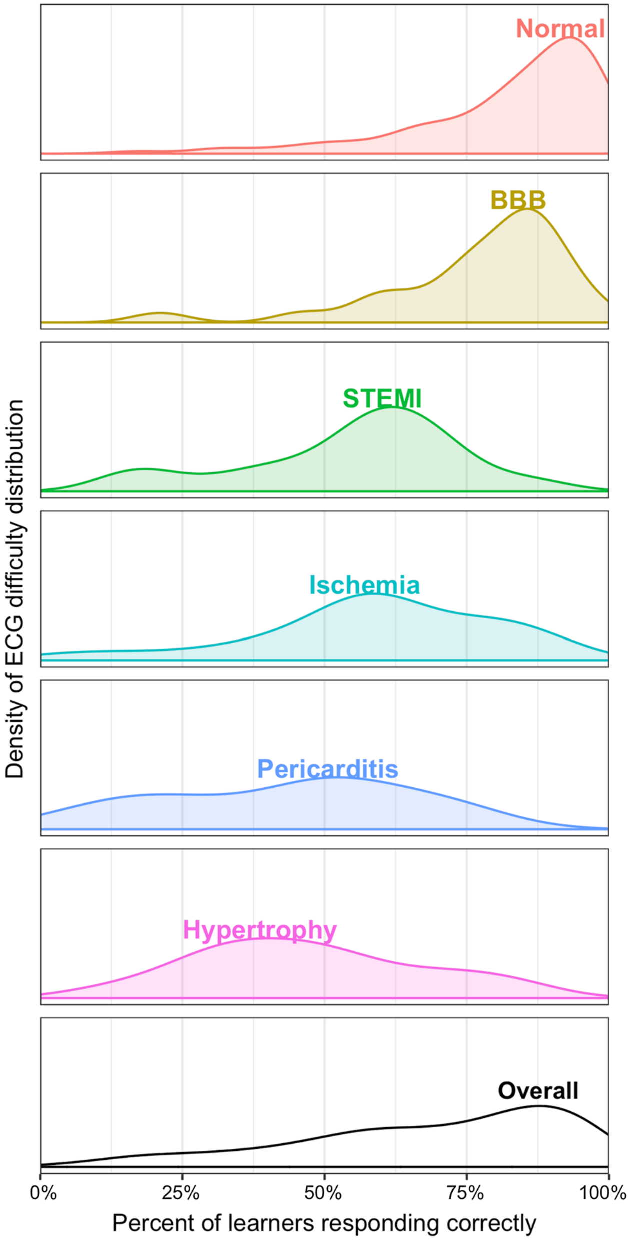


Fig. 3 ECG difficulty distribution by diagnostic category. Depicts the proportion of learners responding correctly to each ECG, plotted against the proportion of ECGs with this percentage of correct responses within each diagnostic category. The bottom panel ('Overall') plots the distribution of percent correct across all ECGs in all diagnostic categories

learning. As demonstrated in Fig. 2b, while learners are in the incremental phase when the outcome measure is diagnostic accuracy, time per ECG is still rapidly decreasing. Thus learners are still learning but the 'rewards' are in time per ECG rather than

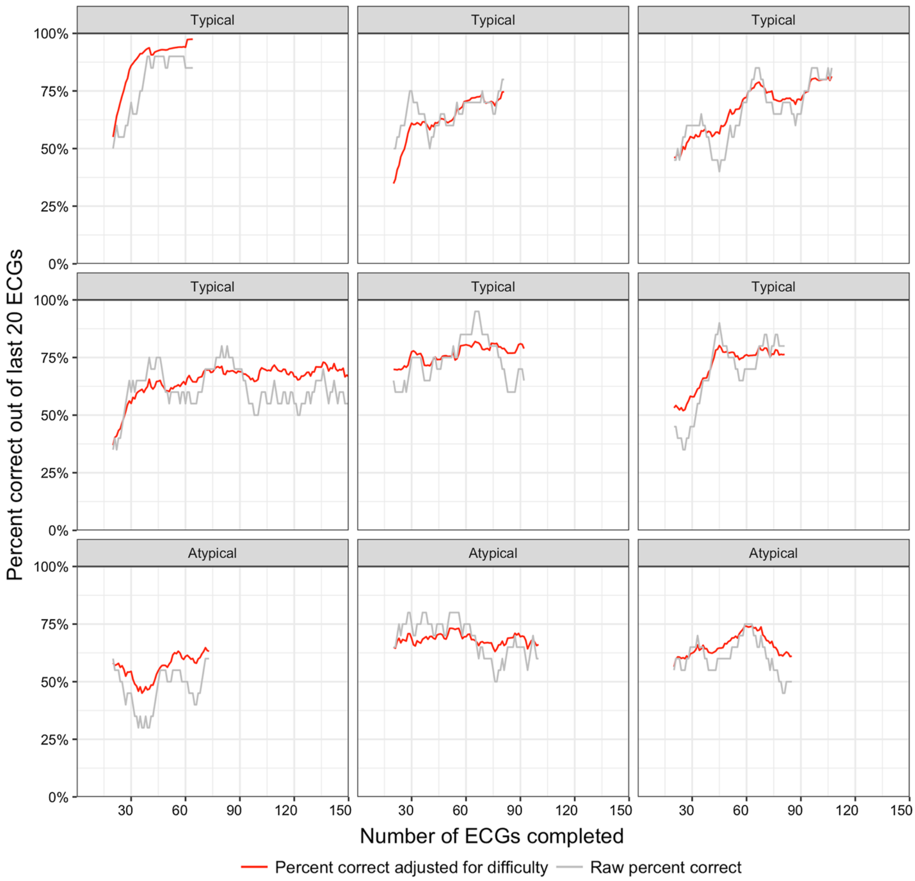


Fig. 4 Individual learning curves for select medical student participants. x-axis = number of ECGs, y-axis = moving average diagnostic accuracy on last 20 ECGs either raw or adjusted for ECG difficulty. Note that significant variability in the shape of the learning curve is visible for the early part of the learning curves (within the first 40 ECGs) but the curves generally settle as more ECGs are interpreted. Of the 63 medical student participants, 39 demonstrated typical learning curves and 24 demonstrated atypical learning curves. (Color figure online)

markedly improved diagnostic accuracy (as occurs during the efficient slope phase of the learning curve).

Examining the proportion of atypical individual learning curves may add further support to the Extrapolation inference by determining whether each learner has the expected learning curve. If a high proportion of learners are not following this curve, then learning is not occurring as predicted. The kappa for two raters agreeing on typical versus atypical learning curve patterns by visual inspection of the 63 medical students' learning curves was 0.89. We found 62% (39/63) had a typical learning curve pattern and 38% (24/63) had an atypical pattern which is consistent with expectations (Pusic et al. 2011). Representative examples of typical and atypical learning curve patterns are presented in Fig. 4. Comparison of the series of nested hierarchical regression models demonstrated that the model that takes into account random variation in slope between individuals has higher explanatory power then the model where all learners are predicted to have the same slope (i.e. slope

Table 2 Log odds ratio and standard error per ECG diagnostic category

ECG diagnostic category (N = number of ECGs = 293)	Learning slope (log odds ratio)	Standard error of learning slope (log odds ratio)
Normal (N = 114)	.28	.03
Bundle branch block (N = 54)	.28	.05
Pericarditis (N = 22)	.13	.06
Ischemia (N = 31)	.12	.05
Hypertrophy (N = 39)	.03	.04
ST elevation myocardial infarction (N = 33)	− .04	.05

The table presents the slope on the log sequence term for each ECG diagnosis, including the reference normal ECGs. The log odds of correctly diagnosing an ECG is based on: the log of the sequence term, dummy indicators for each of the ECG diagnostic categories, and interaction terms between these two terms

from the group-level learning curve) (Supplementary Material). We did not have data on which to fully evaluate the clinical significance of these potential model improvements.

Discussion

In the current study, we provide a full description of learning curves for ECG interpretation skills, using an on-line ECG program. Compared to previous studies of ECG instruction, our learning system focuses on deliberate and mixed practice of ECG interpretation with immediate feedback and uses repeated assessment as opposed to a single post-test (Fent et al. 2015; Chudgar et al. 2016; Rourke et al. 2018). In a recent article on learning curves in health professions' education, we argued for the need for educators to construct the validity argument to support the use of a learning curve for a particular assessment (Pusic et al. 2015a). In the current study, we provide an example of the articulation and examination of the preliminary validity argument that could support a repeated, learning curve-based assessment approach for ECG interpretation. While the data presented are specific to our learning system, the principles of the validity argument are relevant to similar repeated assessment approaches based on mixed practice, deliberate practice and immediate feedback.

Front-line educators are interested in the question "How well is each learner learning"? To answer this, we must have confidence that the assessment approach supports the educational decisions resulting from the assessment. Developing and examining the preliminary validity argument for the repeated assessment of ECG competence within the current study reveals a partially supported argument. Scoring of a learner's response as correct or incorrect is supported by a high level of expert agreement on the correct diagnosis for the ECG. We have support for the sampling aspect of Generalization as ECGs within the program were downloaded from a patient database, and thus reflect a spectrum of real-world ECGs with a representative mix of normal and abnormal (Pusic et al. 2012a; Boutis et al. 2016). The variability of difficulty within and across ECG diagnoses is consistent with prior studies (Hartman et al. 2016; Jablonover et al. 2014). However, limitations in the Generalization inference require further study as we were unable to demonstrate reliability as reflected in more narrow confidence intervals (i.e., increasing consistency) of sequential points along the group learning curves across learners as more ECGs are completed due in part to low sample size of learners completing large numbers of ECGs. We were also unable to undertake generalizability analyses. On the other hand, the slope of the group learning curve was statistically significant, indicating that the majority of learners were learning in this context. Extrapolation from performance on the ECG set to performance on other ECGs is supported by the group learning curves demonstrating expert–novice differences in performance while acknowledging that these differences provide weak evidence (Cook 2015). The performance of the cardiology fellows was higher than the 58% accuracy reported in a previous study of similar participants, suggesting our fellows did represent an 'expert' group (Sibbald et al. 2014). In addition, the proportion of typical to atypical learning curves as identified by visual inspection is similar to prior research (Pusic et al. 2011) and the comparison of the nested hierarchical models indicates that use of a random slope-random intercept model (i.e. individual as opposed to group learning curve data) has the best explanatory power. Implications evidence is required to address issues such as learner engagement and the impact of the assessment on the learner and on the patients they care for.

Limitations and strengths

The weakest inference in the validity argument falls under Generalization. Regarding factor structure, it will be important for future research to determine if there is a metric or set of metrics other than "overall diagnostic accuracy" that more fully captures the development of ECG interpretation skills, particularly for novice learners. For reliability, it will be necessary to formulate a suitable psychometric model for the inferences proposed here, and then to structure data collection to systematically sample the relevant facets of measurement that might contribute to measurement error.

Additional limitations are that the ECG diagnoses did not include arrhythmias, and thus we have not included the full breadth of ECGs that clinicians need to interpret. Broad generalization from our study results should be limited, as in order to generate the learning curves, participants had to interpret at least 60 ECGs and our sample size became smaller as the number of ECGs completed increased. While these small numbers may reflect issues with learner engagement, we suspect they reflect the course context in which the learning system was used, where medical students and residents were instructed to complete a minimum of 30 ECGs. It remains to be seen what level of engagement would occur if learners were asked to achieve a certain level of competence, or were left to engage naturally with the system in the absence of any specific number or standard to be attained. If learning curves model a fundamental phenomenon of human learning (for skills that are amenable to deliberate practice and feedback), then some of the atypical learning curves may reflect a lack of engagement. Future research using qualitative methods to explore the aetiology of atypical curves is required. In addition, we had unequal sized groups for the different levels of learners. Our participants are likely self-selected to be those who are motivated to succeed at ECG interpretation and are all from a single institution. Furthermore, our estimates of item difficulty assume independence across ECGs, which may not hold true in an environment where feedback is given with each ECG (Wainer and Mislevy 2000). We lack evidence supporting the Implications inference.

There are several strengths to the current study. The underlying approach to ECG practice within the on-line system is evidence-based, grounded in the established principles of mixed ECG practice (Hatala et al. 2003) and deliberate practice with immediate feedback

(Ericsson et al. 1993). We have demonstrated the generalizability of the learning curve approach by recruiting learners across varying levels of ECG expertise. Most importantly, we have developed a preliminary validity argument that could be built upon to support a repeated assessment approach for ECG interpretation skills. We emphasize the preliminary nature of this argument and invite the education community to contribute with productive discussion as to how the argument should evolve. The learning curve approach holds the potential to contribute to the assessment of ECG competency, as the data can inform discussion of where to set competency standards. By using a computer-based system that captures multiple metrics, we have demonstrated the ability to continuously monitor the validity argument supporting this repeated assessment.

Implications of the current study

Using a repeated assessment approach, rather than administering a test at a single moment in time, provides an educator and learner with a more robust understanding of where the learner is at in their learning and where they need to go. Fully understanding and responding to where the learner is at by examining individual learning curves (particularly when the curve is atypical) will require discussion between an educator and a learner, as both bring their perspectives on what is happening to bear on their interpretation of the learning curve. With the current study, we have demonstrated the group learning curves for ECG competency for three levels of learners at a single institution. In order to be used more broadly, more learners across multiple institutions will be required. Once the group learning curves are even more robustly established, issues such as how the learner compares with learners at a similar or higher level, or how many ECGs the learner will need to interpret to achieve a certain level of competence, can be addressed. The nested hierarchical models could ultimately be incorporated in a predictive model in order to inform these educational inferences. Having demonstrated that typical learning curves are achievable within this learning platform, an educator could have a learner engage with the ECG learning system and plot their individual learning curve. Examining the individual learning curves, as shown in Fig. 4, allows the educator and learner to understand where the learner is at present (i.e. are they in the efficient early phase? Are they at the incremental phase but still making gains in terms of decreasing the time per ECG? Are they competent at the level of a student, resident or fellow?). Based on this assessment, the educator and learner can collaboratively decide if intervention is necessary and assess any interventions through their impact on the learning curve.

Implementing a repeated assessment approach using a computer-based system that captures both process (e.g. time per page) and outcome (e.g. diagnostic accuracy) metrics with minimal intrusion or cost highlights the potential contribution of a Big Data/Learning Analytics approach (Pecaric et al. 2017). Validity evidence can be collected as a real-time, ongoing process with minimal intrusiveness or cost. As the computer system records multiple metrics simultaneously for large numbers of learners, the data can be analyzed at both the individual learner and the system level. Further, this data can be dynamically incorporated into learning curve models along with features of the item under consideration, such as its difficulty, to allow the learner to benefit from predictive modeling.

For educational researchers, the current study provides fertile ground for future studies. Replication with other learners, at other institutions and with other learning systems based on the principles of mixed and deliberate practice, would help gauge the generalizability of this repeated assessment approach. More research is needed in understanding how learners attain ECG interpretation skills, the significance of atypical learning curves, what concepts are encapsulated in ECG diagnostic accuracy, and whether there are other instructional design features that would yield further learning gains, such as spaced practice or variations on retrieval practice or cues to facilitate dual processing. Furthermore, we need to extend the repeated assessments such that we can subsequently see the forgetting curve which would inform interventions for maintenance of competence (Pusic et al. 2012b).

Validity investigation is an ongoing, iterative process. As learning curve-based assessment systems come to be used more broadly, we will be able to ask and answer new questions and we will be particularly better positioned to evaluate Implications evidence (Cook and Lineberry 2016). While the psychological underpinnings and statistical modelling of group-level learning curves have been fairly well-established both here and in other studies, the analyses of individual learning curves require deeper study (Pusic et al. 2016). The current study provides preliminary insight into how educators interpret learning curves as typical versus atypical; a more complete model of likely interpretations and resulting educational decisions would be very useful going forward.

Conclusion

Our study outlines a validity investigation of a learning-curve-based, assessment for learning, system for ECG interpretation. Our findings generally support foundational inferences in the use of the approach while pointing to rich potential research and development work going forward. In the era of competency-based education, learning curve approaches have the potential to contribute to the discussion around competency standards. Repeated assessment approaches such as those modelled through learning curves also have the potential to provide learners with information as to whether they are 'on-track' and should continue their practice as-is, or whether they require consultation with an educator for additional guidance. Learning curves hold promise for providing educators and researchers, in collaboration with learners, with deeper insights into how well learners are learning.

Funding This work was supported, in part, by a Royal College of Physicians and Surgeons of Canada Medical Education grant.

References

Ashley, E. A., Raxwal, V. K., & Froelicher, V. F. (2000). The prevalence and prognostic significance of electrocardiographic abnormalities. *Current Problems in Cardiology, 25*(1), 1–72.

Boutis, K., Cano, S., Pecaric, M., Welch-Horan, T. B., Lampl, B., Ruzal-Shapiro, C., et al. (2016). Interpretation difficulty of normal versus abnormal radiographs using a pediatric example. *Canadian Medical Education Journal, 7*(1), e68–e77.

Brennan, R. L. (2001). Multivariate unbalanced designs. In R. L. Brennan (Ed.), *Generalizability theory* (pp. 384–387). New York: Springer.

Chudgar, S. M., Engle, D. L., O'Connor, Grochowski C., & Gagliardi, J. P. (2016). Teaching crucial skills: An electrocardiogram teaching module for medical students. *Journal of Electrocardiology, 49*(4), 490–495.

Cook, D. A. (2015). Much ado about differences: Why expert-novice comparisons add little to the validity argument. *Advances in Health Sciences Education, 20*(3), 829–834.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education, 49*(6), 560–575.

Cook, D. A., & Lineberry, M. (2016). Consequences validity evidence: Evaluating the impact of educational assessments. *Academic Medicine, 91*(6), 785–795.

De Bacquer, D., De Backer, G., Kornitzer, M., & Blackburn, H. (1998). Prognostic value of ECG findings for total, cardiovascular disease, and coronary heart disease death in men and women. *Heart, 80*(6), 570–577.

Ericsson, K. A. (2015). Acquisition and maintenance of medical expertise. *Academic Medicine, 90*(11), 1471–1486.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

Fent, G., Gosai, J., & Purva, M. (2015). Teaching the interpretation of electrocardiograms: Which method is best? *Journal of Electrocardiology, 48*(2), 190–193.

Genders, T., Spronk, S., Stijnen, T., & Steyerberg, E. W. (2012). Methods for calculating sensitivity and specificity of clustered data: A tutorial. *Radiology, 265,* 910–916.

Guglin, M. E., & Thatai, D. (2006). Common errors in computer electrocardiogram interpretation. *International Journal of Cardiology, 106*(2), 232–237.

Hartman, N. D., Wheaton, N. B., Williamson, K., Quattromani, E. N., Branzetti, J. B., & Aldeen, A. Z. (2016). A novel tool for assessment of emergency medicine resident skill in determining diagnosis and management for emergent electrocardiograms: A multicenter study. *Journal of Emergency Medicine, 51*(6), 697–704.

Hatala, R. M., Brooks, L. R., & Norman, G. R. (2003). Practice makes perfect: the critical role of mixed practice in the acquisition of ECG interpretation skills. *Advances in Health Sciences Education: Theory and Practice, 8*(1), 17–26.

Jablonover, R. S., Lundberg, E., Zhang, Y., & Stagnaro-Green, A. (2014). Competency in electrocardiogram interpretation among graduating medical students. *Teaching and Learning in Medicine, 26*(3), 279–284.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2008). Test-enhanced learning in medical education. *Medical Education, 42*(10), 959–966.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–197.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education: Macmillan Publishing Company.

Pecaric, M., Boutis, K., Beckstead, J., & Pusic, M. (2017). A big data and learning analytics approach to process-level feedback in cognitive simulations. *Academic Medicine, 92*(2), 175–184.

Pusic, M. V., Andrews, J. S., Kessler, D. O., Teng, D. C., Pecaric, M. R., Ruzal-Shapiro, C., et al. (2012a). Prevalence of abnormal cases in an image bank affects the learning of radiograph interpretation. *Medical Education, 46*(3), 289–298.

Pusic, M. V., Boutis, K., Hatala, R., & Cook, D. A. (2015a). Learning curves in health professions education. *Academic Medicine, 90*(8), 1034–1042.

Pusic, M. V., Boutis, K., Pecaric, M. R., Savenkov, O., Beckstead, J. W., & Jaber, M. Y. (2016). A primer on the statistical modeling of learning curves in health professions education. *Advances in Health Sciences Education, 22*(3), 741–759.

Pusic, M. V., Chiaramonte, R., Gladding, S., Andrews, J. S., Pecaric, M. R., & Boutis, K. (2015b). Accuracy of self-monitoring during learning of radiograph interpretation. *Medical Education, 49*(8), 838–846.

Pusic, M. V., Kessler, D., Szyld, D., Kalet, A., Pecaric, M., & Boutis, K. (2012b). Experience curves as an organizing framework for deliberate practice in emergency medicine learning. *Academic Emergency Medicine, 19*(12), 1476–1480.

Pusic, M., Pecaric, M., & Boutis, K. (2011). How much practice is enough? Using learning curves to assess the deliberate practice of radiograph interpretation. *Academic Medicine, 86*(6), 731–736.

Ramsay, C. R., Grant, A. M., Wallace, S. A., Garthwaite, P. H., Monk, A. F., & Russell, I. T. (2001). Statistical assessment of the learning curves of health technologies. *Health Technology Assessment (Winchester, England), 5*(12), 1–79.

Rourke, L., Leong, J., & Chatterly, P. (2018). Conditions-based learning theory as a framework for comparative-effectiveness reviews: A worked example. *Teaching and Learning in Medicine, 16,* 1–9.

Salerno, S. M., Alguire, P. C., & Waxman, H. S. (2003a). Competency in interpretation of 12-lead electrocardiograms: A summary and appraisal of published evidence. *Annals of Internal Medicine, 138*(9), 751–760.

Salerno, S. M., Alguire, P. C., & Waxman, H. S. (2003b). Training and competency evaluation for interpretation of 12-lead electrocardiograms: Recommendations from the American College of Physicians. *Annals of Internal Medicine, 138,* 747–750.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011a). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher, 33*(10), 783–797.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011b). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher, 33*(6), 478–485.

Shah, A. P., & Rubin, S. A. (2007). Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *Journal of Electrocardiology, 40*(5), 385–390.

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Sibbald, M., Davies, E. G., Dorian, P., & Yu, E. H. C. (2014). Electrocardiographic interpretation skills of cardiology residents: Are they competent? *Canadian Journal of Cardiology, 30*(12), 1721–1724.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing* (pp. 63–68). New Jersey: Lawrence Erlbaum & Associates.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (pp. 81–124). Amsterdam: Elsevier.

Affiliations

Rose Hatala[1] · Jacqueline Gutman[2] · Matthew Lineberry[3] · Marc Triola[2] · Martin Pusic[2,4]

[1] Department of Medicine, St. Paul's Hospital, University of British Columbia, Suite 5907, Burrard Bldg, 1081 Burrard St, Vancouver, BC V6Z 1Y6, Canada

[2] Institute for Innovations in Medical Education, New York University School of Medicine, New York, NY, USA

[3] Zamierowski Institute for Experiential Learning, University of Kansas Medical Center and Health System, Kansas City, KS, USA

[4] Ronald O. Perelman Department of Emergency Medicine, New York University School of Medicine, New York, NY, USA