BMJ Open

Protocol for development of an assessment tool for competency of ECG interpretation: expert consensus by the RAND/UCLA appropriateness method and cross-sectional testing using multidimensional item response theory

Shinji Inaba,1 Kazumichi Yamamoto,2,3 Tomohiro Kaga,4 Muhammad Wannous,3,5 Masatsugu Sakata,2 Osamu Yamaguchi,6 Toshi A Furukawa2

ABSTRACT

Introduction Although the ECG is an important diagnostic tool in medical practice, the competency of ECG interpretation is considered to be poor. Diagnostic inaccuracy involving the misinterpretation of ECG can lead to inappropriate medical judgements and cause negative clinical outcomes, unnecessary medical testing and even fatalities. Despite the importance of assessing ECG interpretation skills, there is currently no established universal, standardised assessment tool for ECG interpretation. The current study seeks to (1) develop a set of items (ECG questions) for estimating competency of ECG interpretation by medical personnel by consensus among expert panels following a process based on the RAND/UCLA Appropriateness Method (RAM) and (2) analyse item parameters and multidimensional latent factors of the test set to develop an assessment tool.

Methods and analysis This study will be conducted in two steps: (1) selection of question items for ECG interpretation assessment by expert panels via a consensus process following RAM and (2) cross-sectional, web-based testing using a set of ECG questions. A multidisciplinary panel of experts will evaluate the answers and appropriateness and select 50 questions as the next step. Based on data collected from a predicted sample size of 438 test participants recruited from physicians, nurses, medical and nursing students, and other healthcare professionals, we plan to statistically analyse item parameters and participant performance using multidimensional item response theory. Additionally, we will attempt to detect possible latent factors in the competency of ECG interpretation. A test set of question items for ECG interpretation will be proposed on the basis of the extracted parameters.

Ethics and dissemination The protocol of this study was approved by the Institutional Review Board of Ehime University Graduate School of Medicine (IRB number: 2209008). We will obtain informed consent from all participants. The findings will be submitted for publication in peer-reviewed journals.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This protocol describes a study for the development of an assessment tool for the competency of ECG interpretation using expert panel consensus and multidimensional item response theory (IRT).
⇒ This study proposes a clear procedure for creating a test set to assess competency in ECG interpretation based on prespecified criteria and consensus of a panel of experts.
⇒ Multidimensional IRT provides objective and quantitative parameters such as item difficulty, item discrimination and potential latent factors for assessment tools.
⇒ Item parameters depend on the response pattern of participants and the results could potentially be influenced by the distribution of responders’ competency.
⇒ The absence of a generally accepted gold standard for assessing competency in ECG interpretation makes it difficult to confirm criteria validity related to external benchmarks.

INTRODUCTION

Although the ECG is an important diagnostic tool in medical practice, the competency of ECG interpretation is considered to be poor, particularly among undergraduates and postgraduates.1 According to a recent review article, even specialists in cardiology correctly identify ECG abnormalities in only 53%–96% of cases.2 Diagnostic inaccuracy caused by the misinterpretation of ECG can lead to inappropriate clinical medical decision-making and can cause negative clinical outcomes, unnecessary medical testing and even fatalities.3

ECG interpretation involves the integration of multiple domains of knowledge, such as
anatomy, physiology, visual wave recognition and medical reasoning. Despite its complexity and difficulty, there is no adequate evidence to support specific training, methods for learning ECG interpretation, standardised assessment tools to evaluate ECG reading competency by medical personnel or approaches for the maintenance of interpretation skill.\(^7\)\(^8\)

Despite the importance of assessments of ECG interpretation skill, in a recent systematic review, Cook \textit{et al} reported that there is currently no established universal, standardised assessment tool for ECG interpretation for the broader community.\(^6\) The American Board of Internal Medicine annually conducts a cardiovascular disease certification exam that includes an ECG and diagnostic imaging component and publishes the criteria for its generation (https://www.abim.org/Media/ksmt11c/cardiovascular-disease.pdf). The process for creating this exam is systematically reviewed and updated annually by a multidisciplinary panel of experts. However, the examination is intended more for candidate specialists and is not available to general residents or other medical professionals. In addition, the process relies heavily on the knowledge and experience of the specialist, and an evaluation tool based on an objective process and intended for a wider range of trainees is awaited.

Methods for evaluating the validity of assessment tools have been extensively investigated and developed in psychological and educational disciplines.\(^7\)\(^8\) In 1999, the American Educational Research Association, American Psychological Association and National Council on Measurement in Education proposed standards for educational and psychological testing to integrate concepts of validity and reliability.\(^9\) This report stated that ‘validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests’ and proposed the ‘five evidence sources’ validity framework, which examined evidence regarding content, response process, internal structure, relationships with other variables and consequences.

To develop summative assessments, the process of item generation (ECG selection), correct answers and scoring rubrics are particularly important.\(^6\) Content evidence refers to the ‘relationship between a test’s content and the construct it is intended to measure’.\(^9\) In practice, content evidence is assessed by estimating how the assessment tools were created, and requires a detailed description of the development of tools.\(^10\) In medical research, there are several methods to develop a consensus regarding the collective opinion of experts.\(^11\) Of these, the RAND/UCLA Appropriateness Method (RAM), also called the modified Delphi method, was developed to assess the appropriateness of medical interventions. As the name indicates, this method was first developed to assess how appropriate or necessary medical procedures are for a certain medical condition. However, this method is also used to seek consensus among expert panels.\(^12\)\(^13\) RAM consists of three primary components: literature review and preparation of a list of topics, individual questionnaires for each expert panellist to answer, followed by a meeting with all the panellists together to reach a consensus. Clear descriptions are important for content validity.

Among the five sources of evidence, internal structure is also an important factor. Traditionally, to assess the internal structure of questionnaires or assessment tools, reliability and factor analysis have generally been used, on the basis of classical test theory (CTT).\(^14\) Item response theory (IRT) has been increasingly used in recent years, particularly in psychological and educational research.\(^15\) Compared with CTT, which depends on the linear relationship between scale items and its latent factors, IRT models the probability of a response pattern of items on the basis of a non-linear model. This relationship between CTT and IRT is like that of linear and logistic regression, although IRT is not a regression model in the strict sense.\(^16\)

IRT embodies a wide variety of statistical models to express item characteristics of psychometric tools by estimating an individual’s response pattern and, in most cases, item discrimination and threshold are the parameters of interest. For example, in the case of a dichotomously scored assessment tool for estimating the competency of ECG interpretation, item discrimination refers to the degree to which the item discriminates the response (correct or incorrect) along the underlying continuum of a latent trait (eg, a certain responder’s competency of ECG interpretation) and threshold parameter refers to the point at which the responder answered the item correctly with a probability of 0.5. On the basis of these parameters, an individual’s latent trait (eg, competency, in this case) is expressed as the point of standing on the continuum of latent traits.\(^16\)

IRT has several advantages over CTT. One of these advantages is that IRT estimates item parameters independently from the sample set because the estimation is based on the relative relationship among the response patterns, and an individual’s latent traits are not required.\(^17\) Accordingly, the latent trait of each individual is also independent from the administered items because latent traits are calculated from the item parameters of each item. This characteristic is beneficial, particularly for educational tests, for which a different set of items is required each time.

Despite these advantages, IRT involves one important limitation. Because traditional IRT is based on the assumption of unidimensionality, it has not been used to estimate multidimensional assessment tools. Given this limitation, multidimensional IRT (MIRT) has been developed and used as an alternative to factor analysis such as exploratory factor analysis and confirmatory factor analysis on the basis of CTT.\(^15\)\(^16\)\(^18\)

To the best of our knowledge, no previous studies have examined multidimensionality in competency of ECG interpretation. In most previous research about assessment of ECG interpretation competency, the classifications of ECG questions were based on the pathophysiology
of the underlying disease, such as ischaemic, arrhythmic, structural and metabolic. However, there is no evidence that aetiological classification reflects latent factors of ECG interpretation.

The purpose of the current study is (1) to develop a set of items (ECG questions) to estimate competency of ECG interpretation by consensus of expert panels following the process based on RAM and (2) to analyse item parameters and multidimensional latent factors of the test set to develop an assessment tool.

METHODS AND ANALYSIS
The protocol of this study was approved by the Institutional Review Board of Ehime University Graduate School of Medicine (IRB number: 2209008). We will obtain informed consent from all participants before their participation in the assessment. All of the procedures will be conducted in accordance with the Declaration of Helsinki.

Study design
This study will be conducted in two steps: (1) selection of question items for ECG interpretation by expert panels via a consensus process following RAM and (2) cross-sectional, web-based testing using the above-mentioned set of ECG questions. On the basis of the collected data from the test, item and person parameters will be statistically analysed and a test set of question items for ECG interpretation will be proposed.

Step 1: selection of question items for ECG interpretation
Settings and design
This step will be conducted by expert panels following RAM. The consensus process will be conducted in three phases: a preparatory phase (round 0), an online-test phase (round 1), and a face-to-face expert panel meeting (round 2).

Participants
According to the RAM, the number of expert panels should be between 7 and 15, with 9 in particular being traditionally employed. We will recruit nine expert panels who deal with ECG as specialists or in daily practice, in accord with the recommendations of RAM. Ehime University has a medical network system in Ehime Prefecture, and panellists will be recruited from hospitals, the school of medicine of Ehime University, and fire stations in Ehime Prefecture. The titles of the selected panel members will be as follows: two cardiologists, two cardiologists subspecialised in arrhythmia, one cardiologist subspecialised in adult congenital heart disease, one emergency physician, one paramedic, one clinical laboratory technician and one registered nurse who is currently a medical student. All panellists are board-certified cardiologists, ECG decoders or have at least 10 years of experience in their specialty.

Procedures
The first phase is the preparatory phase. After preparing a summary of recent literature regarding the assessment of ECG interpretation, a document about RAM, and a document explaining how the consensus process is conducted, a preparatory meeting (round 0) will be conducted to discuss the aim and process of this study as well as additional topics to be assessed, such as candidates for latent factors regarding competency of ECG interpretation, the method for rating the appropriateness of each question as an assessment tool, and the number of item questions of ECG to be assessed by the expert panels.

On the basis of the results of round 0, a total of 100 candidate question items for ECG interpretation will be selected. At Ehime University, the department of cardiology periodically offers structured, face-to-face lecture courses via an online video conference system. One term consists of a 40 min lecture, twice a week for 5 weeks, with theoretical explanations followed by exercises using graphical ECG waves with a short clinical presentation. The 100 candidate question items for ECG interpretation will be selected from the item pool of more than 500 items from this lecture course according to the Minnesota code to cover as wide a range as possible by two of the authors (SI and JA) who mainly lead this lecture course.

In the next phase, the panellists will be asked to answer individually for a total of 100 candidate questions in an online system developed for this study (round 1). Each question contains a short clinical presentation and a graphic picture of ECG waves with five response choices, together with a free comment section to describe the rationale for the diagnosis. Panellists will be asked to answer regarding two types of topics (ie, diagnosis and latent factor). To investigate the multidimensionality of the competency for ECG interpretation, panellists will be asked to answer questions about a possible latent factor, to which each item question could belong, from the list of choices based on the results of Round 0. Potential latent factors for ECG interpretation have not been systematically studied and are often categorised by physiological factors such as ischaemia or arrhythmia. In round 0 of this panel meeting, we will discuss candidate potential latent factors, including the possibility of other factors. In addition to the answers regarding ECG interpretation, the panellists will also be asked to answer about the appropriateness of each item as an assessment tool, taking into account the clarity of the clinical presentation, the quality of the ECG wave, and the discrimination of the choices of answers using a nine-point Likert scale, ranging from 1 (completely irrelevant) to 9 (extremely relevant). The panellists will be asked to provide free comments if they have any suggestions or rationale for their decision.

In the final phase, a face-to-face expert panel meeting will be held to review on the basis of the results of round 1 (round 2). The meeting will be chaired by a physician specialised in clinical epidemiology who will not have been involved in round 1 (KY). Before the meeting, each panellist will have received the results of round 1
including their own answer for each item question with the summary of results from all of the panellists. Panellists will be blinded to all of the answers of the other panellists. Regarding the answer and latent factor for each item question, we will estimate the degree of agreement using Fleiss’ kappa. The kappa estimate (k) is considered acceptable if k is more than 0.6 and any inconsistency will prompt discussion to investigate possible reasons. The degree of agreement for each question is determined by expressing the percentage of agreement with the prepared answers. Regarding the appropriateness of each item question, the median and distribution of the answers as well as the IQR will be reported. Item questions with a median score between 7 and 9 with IQR≤2 are considered appropriate for the assessment tool, and item questions with a median of 1 and 3 with IQR>2 are considered inappropriate. If the median value is between 4 and 6 or IQR is more than 2, the expert panels will discuss the reason for the inconsistency of the results. On the basis of the results of three phases, a total of 50 item questions with answers and possible latent factors will be prepared as an assessment tool for the next step.

Step 2: implementation of online assessment

Design and setting
This part of the study is a cross-sectional, online assessment. The assessment will be administered through a web-based system developed for this study.

Participants
The target population of this phase will not be limited to medical physicians and undergraduates but will include various individuals who are likely to evaluate ECG on a regular basis, including cardiologists, non-cardiology physicians, residents, undergraduates, nurses, nursing students, medical engineers, paramedics and medical clerks. Ehime University has a medical network system in Ehime Prefecture, and participants will be recruited from hospitals, the school of medicine of Ehime University, the school of nursing, and fire stations in Ehime Prefecture. Inclusion criteria for participants will be as follows: (1) aged 18 years old or more; (2) individuals who have undergone a systematic course about the interpretation of ECG, (educational courses such as medical, health, and nursing schools, as well as systematic courses as post-graduation continuing education), (3) individuals who have engaged in medical-related activities in the previous 2 years; (3) individuals who understand Japanese sufficiently to answer questions about the assessment tool. Exclusion criteria will be as follows: (1) individuals who have never undergone a systematic lecture or course about the interpretation of ECG; (2) individuals who have had no opportunities to evaluate ECG on a regular basis. As an incentive, an Amazon gift card worth ¥2000 (ca. US$15) will be awarded to all participants.

Process of assessment
Interested individuals will register their intent to participate in the study through a Google Form web page that requests the respondent’s name, age, academic title, current job title, experience of systematic lecture or course for ECG interpretation and email address. If the candidate fulfils the inclusion criteria, an invitation with ID and password will be sent by email. After logging in to the system, participants will be asked to accept the conditions of participation in the study. At the beginning of the assessment, participants will be required to answer a questionnaire regarding demographic information. A total of 50 item questions will then be administered in a random order with a time limit of 90 min. At the end of the assessment, the number of correct answers among 50 questions will be shown without further detail.

Measurements
As demographic information, the following data will be collected: age, sex, academic title, current job title, years of experience after education about ECG interpretation, frequency of ECG interpretation (low, intermediate, high). The frequency of ECG readings was defined as ‘low’ for less than once a week, ‘high’ for every working day and ‘intermediate’ for somewhere in between. Regarding the item questions about ECG interpretation, participants will be asked to choose one correct answer from five choices.

Statistical analysis
Sample size
There is no clear standard for the minimum sample size for IRT, and previous studies recommended a wide range of possible numbers, from 50 to 100 for one-parameter models to 2000 for MIRT, because small sample sizes can lead to unstable item parameters. In a simulation study using multidimensional graded response models, Jiang et al reported that a sample size of 500 participants provided accurate parameter estimates in most scenarios, and increasing the sample size did not substantially improve the accuracy of parameter estimation. In another study conducted to assess latent factor structure using MIRT, the authors recommended that a sample of at least 350 participants was required to avoid instability of parameter estimates. Taking into account that this study is a binary response model, the required sample size was calculated on the basis of these findings, including dropouts and incomplete data. To obtain at least 350 complete sets of response data with a dropout or incomplete response rate of 20%, we estimate that 438 participants will be required.

Model selection, estimation and model fit
The data analysis will consist of two parts. First, to investigate dimensionality, exploratory factor analysis will be performed using IRT. An exploratory IRT model will be implemented using the following approach: (A) transformation of dichotomous response using quasipolychoric correlation; (B) parameters estimation using
the Metropolis-Hastings Robbins-Monro algorithm\textsuperscript{23}, (C) factor rotation using an oblique rotation (geomin) and (D) factor loadings of at least 0.40 are required for factor inclusion.\textsuperscript{24} Several possible models are then estimated according to the number of dimensions, on the basis of consensus via the panel discussion, and model-data fit are compared among models using the –2LL statistic, Akaike’s information criterion (AIC),\textsuperscript{25} Bayesian information criterion (BIC)\textsuperscript{26} and root mean square error of approximation (RMSEA).\textsuperscript{15} Regarding RMSEA, a value of less than 0.08 is preferred for model-data fit. On the basis of model fit indices and a physiological theorem, the number of dimensions will be determined for further analysis. For the items, the local dependence hypothesis will be examined by detecting residual covariation between items\textsuperscript{27} and Cramer’s V<0.2 is deemed satisfactory.

Second, to assess multidimensionality, IRT analysis will be conducted on the basis of a 3PL model with a guessing value of 0.2 because all of the responses are binary (ie, correct/incorrect) from five choices. Parameters will be estimated using the Gauss-Hermite quadrature method used in traditional EM estimation.\textsuperscript{28} Tested models include unidimensional and bifactor models, and model fit will be assessed using the –2LL statistic, AIC, BIC and RMSEA.\textsuperscript{15}

**Test/item/person diagnosis**

Item fit will be assessed using Zh statistics\textsuperscript{29} for all of the items; Zh statistics >0 indicate a better fit than expected, whereas values of less than 0 indicate a worse fit than expected. Item discrimination and item intercept parameters for each item will then be reported together with the multivariate discrimination and intercepts for determining an item’s overall utility across factors.\textsuperscript{30}

**Latent trait (θ) prediction**

Latent trait (θ) of each respondent is estimated using maximum a priori estimation, a Bayesian method using an iterative procedure with a normal prior distribution with a mean of 0 and SD of 1.\textsuperscript{31}

**Statistical software**

We will use R V.4.2.1 (R Foundation for Statistical Computing, Vienna, Austria) for all of the analyses. We will use the ‘mirt’ package for the MIRT.\textsuperscript{26}

**Patient and public involvement**

The research questions and outcome measures were developed to address the concerns and interests of healthcare professionals to ensure that the study results are applicable to them. Healthcare providers were involved in the design of the study through feedback and input obtained from the consultation. The recruitment strategy will target healthcare professionals through community outreach, advertising in public spaces, and online platforms. Study results will be distributed to study participants in accessible and understandable ways, such as journal articles.

**Study status and timeline**

At the time of preparation of this protocol for publication, we are preparing to begin the first phase of the study (expert panel meeting). We anticipate 3 months to complete the expert panel meeting to select the final set of test questions. Recruitment of participants for the cross-sectional study is expected to take about 18 months.

**ETHICS AND DISSEMINATION**

The protocol of this study was approved by the Institutional Review Board of Ehime University Graduate School of Medicine (IRB number: 2209008). We will obtain the informed consent from all of the participants before the participation in the assessment. The findings will be submitted for publication in peer-reviewed and open access journals.

**DISCUSSION**

We have presented a protocol for developing a tool to assess the competency of ECG interpretation by medical personnel of various backgrounds and training levels. To the best of our knowledge, this is the first reported study protocol for developing an assessment tool for the competency of ECG interpretation using IRT. This study will contribute to the development of a statistics-based assessment tool using parameters extracted from IRT analysis, not only for evaluating the competency of responders, but also for selecting adequate questions as a test set. Once we are able to obtain a test set of questions on the basis of these parameters, we will be able to standardise the generation of a test set for the assessment tool, which is developed individually by experts in each setting.

Moreover, we will be able to develop a computerised adaptive testing (CAT) system on the basis of these parameters, which will be extracted from this study.\textsuperscript{17} CAT selects each question based on individual participants’ response patterns using item parameters until the predicted latent trait of responders is considered statistically stable. This process will enable us to automate the generation of a test set and assess the competency of responders with fewer questions, compared with a classic fixed test set.\textsuperscript{17} This study will provide a possible candidate test set with item parameters.

There are several limitations involved in this study protocol that should be considered. First, the participants in the study could be biased because item parameters depend on the response pattern of participants and the results could potentially be influenced by the distribution of responders’ competency. We will attempt to recruit a large sample of participants from a broad range of backgrounds to help reduce bias. Second, we will not be able to confirm criterion validity, which is the relationship with external benchmarks, because there is no accepted gold standard for assessing the competency of ECG.
interpretation. Further study will be required in different settings to evaluate the generalisability of the parameters. Regarding future research directions, there are several promising possibilities. Once we develop the first test set of the tool, we can continue to recruit participants to stabilise the parameters. Additionally, we can add more questions to evaluate the parameters on the basis of parameters from existing questions (eg, item linking) which could further stabilise the quality of the assessment tool. As mentioned above, we can also develop a system of CAT on the basis of these parameters. The proposed study protocol aims to develop an evidence-based system to evaluate the competency of ECG interpretation to standardise the assessment method.

Author affiliations
1Department of Cardiology, Pulmonology, Nephrology and Hypertension, Ehime University Graduate School of Medicine, Toon, Japan
2Departments of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine Faculty of Medicine, Kyoto, Japan
3Institute for Airway Disease, Hyogo, Japan
4Ehime University, Matsuyama, Japan
5Department of Computer Information Science, Higher Colleges of Technology, Abu Dhabi, UAE
6Department of Cardiology, Ehime University Graduate School of Medicine, Toon, Japan

Twitter Toshi A Furukawa @Toshi_FRKW

Acknowledgements We are extremely grateful to the members of expert panel, which will comprise the following individuals: Takayuki Nagai, Jun Aono, Haruhiko Higashi, Hiroshi Kawakami, Muneaki Ohshita, Yusuke Akazawa, Tomoki Fujisawa, Yukari Shikano from Ehime University Graduate School of Medicine, and Yasuyuki Watanabe from Imabari City Fire Department. We thank Benjamin Knight, MSc., Inaba S, et al. for the conduct of this study. Contributors All authors contributed to the conception and conduct of this study. Of is the principal investigator of this study protocol. SI, KY and TK designed the overall framework of this study and CS, OF and TAF helped with implementation. MW developed the online system which will be used in this study. SI, KY and TK drafted the manuscript, and all coauthors read it and provided critical comments. All authors read and approved the final manuscript.

Funding This study is funded by industrial seeds support from Ehime University, Ehime, Japan.

Competing interests TAF reports personal fees from Boehringer-Ingelheim, DT Axis, Kyoto University Original, Shinogi and SonNY, and a grant from Shinogi, outside the submitted work. In addition, TAF has patents 2020-548587 and 2022-082495 pending, and intellectual property for Kokoro-app licensed to Mitsubishi-Tanabe. KY and MW have patent 2022-193813 pending outside the submitted work. MS reports personal fees from SONNY outside the submitted work. The Institute for Airway Disease received a system development fee from the Department of Cardiology, Pulmonology, Nephrology and Hypertension, Ehime University Graduate School of Medicine. All other authors have no conflict of interest.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

REFERENCES
30 Reckase MD, McKinley RL. The discriminating power of items that measure more than one dimension. *Appl Psychol Measure* 1991;15:361–73.