Sources of Measurement Error in an ECG Examination: Implications for Performance-Based Assessments

David J. Solomon, Ph.D. and Gary Ferenchick, M.D.

Michigan State University

Running Head – Sources of Measurement Error in an ECG Exam

The paper was presented in poster format at the Clerkship Directors of Internal Medicine Conference, Washington, DC, October 2002.

Corresponding Author:

David J. Solomon, Ph.D.

A-202 E. Fee Hall

Michigan State University

E. Lansing, MI 48823

Phone: (517) 353-2037

Fax: (517) 432-1798

E-mail dsolmon@msu.edu

Abstract

Objective: To assess the sources of measurement error in an electrocardiogram (ECG) interpretation examination given in a third-year internal medicine clerkship.

Design: Three successive generalizability studies were conducted. 1) Multiple faculty rated student responses to a previously administered exam. 2) The rating criteria were revised and study 1 was repeated. 3) The examination was converted into an extended matching format including multiple cases with the same underlying cardiac problem.

Results: The discrepancies among raters (main effects and interactions) were dwarfed by the error associated with case specificity. The largest source of the differences among raters was in rating student errors of commission rather than student errors of omission. Revisions in the rating criteria may have helped increase inter-rater reliability slightly however, due to case specificity, it had little impact on the overall reliability of the exam. The third study indicated the majority of the variability in student performance across cases was in performance across cases within the same type of cardiac problem rather than between different types of cardiac problems.

Conclusions: Case specificity was the overwhelming source of measurement error. The variation among cases came mainly from discrepancies in performance between examples of the same cardiac problem rather than from differences in performance across different types of cardiac problems. This suggests it is necessary to include a large number of cases even if the goal is to assess performance on only a few types of cardiac problems.

Keywords: Electrocardiogram, educational, measurement, generalizability, performance based assessment, reliability

Over the last several decades there has been an increasing realization that the evaluation of medical trainees must include measures that reflect the complex multidimensional activities physicians perform in real-world settings. This has lead to the use of a variety of evaluation techniques under the rubric of performance-based assessment (PBA). The term "performance-based assessment" has been used in a number of ways. For the purpose of this paper we are using a definition provided by Mavis and colleagues (Mavis, Henry, Ogle & Hoppe, 1996). Performance should involve the integration of multiple capabilities; involve the observation of behaviors; be clinically relevant, linked to actual clinical tasks and cued only insofar as clinical tasks are cued in the real world.

PBAs are designed to measure performance in ways that more closely reflect the activities of clinicians in the real world. At the same time, they appear to provide more rigorous measurement than direct observation of trainees in clinical teaching settings such as preceptor ratings of clerks or residents in hospital and clinic rotations. There is a price to pay for theses advantages (Downing, 2002). First, conducting PBAs tends to be complicated, expensive and time consuming. Secondly, as noted by Swanson, Norman and Linn (1995), developing and interpreting these assessments poses a number of daunting challenges.

One of the skills often taught in internal medicine clerkships is the interpretation and use of twelve lead electrocardiograms (ECGs) in the care of the patient. We evaluate students in our third year clerkship on their ability to interpret and apply information from an ECG in a PBA. Students are asked to assess rate, rhythm, axis, measure intervals, provide a diagnosis and to propose initial management steps from a series of ECGs and short patient scenarios. Over a two-year period we conducted a series of studies in order to determine the generalizability and estimate the magnitude of the sources of measurement error in our ECG interpretation

examination. Our goal was to redesign the PBA in ways that increase its reproducibility while minimizing the effort of conducting the exercise for the faculty, staff and students involved. We feel our findings will be relevant to other medical educators who are developing or currently using similar performance-based measures.

Methods

Michigan State University's (MSU) College of Human Medicine (CHM) is a community-integrated medical school. Medical students spend their last two years in one of six community-based campuses distributed across the State of Michigan. Each campus has a community internal medicine clerkship director.

Basic techniques in reading and interpreting 12-lead ECGs are taught as part of our eight-week third-year internal medicine clerkship. The students are evaluated via a performance-based examination where they are required to interpret and provide management recommendations from a series of 12-lead ECG and short patient scenarios. Originally, students responded to each of 10 ECG/patient scenarios by completing a series of short answer questions. The cases included one normal ECG/scenario and nine different cardiac problems including myocardial infarction, hypertrophy, bundle branch and AV nodal blocks, atrial fibrillation and flutter. The PBAs were graded by the community internal medicine clerkship director or their designate in each of the six communities.

We became concerned about the inter-rater reliability of the scoring of the PBA and therefore completed a series of three generalizability studies. In the fall of 2000 we conducted our first generalizability study. This included data from 10 students interpreting six ECGs rated by seven faculty members. The 10 student examinations were selected from previous

administered examinations such that they reflected the observed distribution of scores seen among the students in the clerkship.

The faculty scoring each of the short-answer exams used rating criteria that were previously developed and used for grading the exam. After completing the study, discrepancies in the ratings were discussed by the faculty participants to determine the cause of these discrepancies. In the spring of 2001 we revised rating criteria based on the findings of the first study and completed our second generalizability study, this time with 11 faculty (six of the original raters and five new raters) using 10 cases and five student examinations. The change in design reflected our attempt to gather better data on the variability among raters and cases while keeping the time commitment of the faculty raters under an hour.

The following academic year (2001-2002) we converted the examination to an extended matching format and increased the number of cases included in the examination from 10 to 15. This allowed us to include two cases with the same underlying cardiac problem for four separate cardiac problems as well as including two normal ECGs. At the end of the 2001-2002 academic year, we conducted a third generalizability study using the actual examination data from the all students completing the clerkship during the 2001-2002 academic year. This analysis included 92 students, with two cases nested within four types of cardiac problems and two normal ECGs.

In each of the studies, each case was rated/scored on a 10 point scale. In all three studies, estimates of the variance components and their standard errors were calculated using GENOVA (Crick & Brennan, 1984). Estimates of the generalizabilty of specific measurement designs presented in Figure 1 are based on methods discussed by Brennen (2001). Data from new and repeat raters who participated in study 2 were combined and used for the decision or "D" study presented in Figure 1 along with data from Study 3.

Results

Study 1: The estimates of the variance components and their standard errors are presented in Table 1 along with the results from Study 2. The results of first study indicated the largest source of error was the student by ECG interaction or what is often termed "case specificity." Though the discrepancies among raters (both interactions and main effects) accounted for some of the error in the ratings, it was dwarfed by the error associated with case specificity. The magnitude of the variance components was roughly 0.6 versus 2.4 respectively. The review of the discrepancies among raters indicated major source of the differences was in rating student errors of commission rather than student errors of omission. That is, when students stated something that was clearly wrong rather that when students failed to include the correct response for a question. Based on the findings the rating instructions were modified to more clearly delineate how to score errors of commission.

Study 2: As in the first study, the largest source of error was the student by ECG interaction or case specificity. The variance associated with raters dropped from roughly 0.6 to 0.3, possibly based on the improved rating criteria though the difference was small compared with the standard error. The error variance associated with raters also had little effect on the generalizability of the PBA. The estimates of the variance components as their standard errors for Studies 1 and 2 are contained in Table 1.

Study 3: Since we moved to an extended matching format, the design for this study did not include raters. The design included students, cardiac problems and cases nested within cardiac problem. This allowed us to disentangle the variance associated with types of cardiac problem from the variance associated with individual cases within a particular type of cardiac problem. The estimated variance components and their standard errors are provided in Table 2.

Virtually all the error variance was accounted for by cases within cardiac problem and the interaction of students and the cases within cardiac problem. Almost none of the variance was accounted for by cardiac problem or the interaction of students and cardiac problem (4.5 versus 0.03).

Figure 3 presents a "D study" analysis comparing the generalizability of different combinations of raters (for the short-answer format) and cases.

Conclusions

Our initial reason for embarking on these studies was our concern over the consistency of the scoring the PBA by the different clerkship directors in the six CHM community campuses. We determined from the first study that the major source of the discrepancies in ratings among our community clerkship directors involved errors of commission rather than errors of omission. The results of study 2 suggest revising our rating criteria may have reduced these discrepancies somewhat however it had little impact on the reproducibility of the ratings. The overwhelming source of error was case specificity or the tendency for students to perform inconsistently across the various cases in the examination.

In retrospect the results are not at all surprising given nearly 30 years of research on physician performance has largely result in similar findings, of example (Downing, 2002; Swanson, Norcini, 1989; Van Thiel, Kraan, Van Der Vleuten, 1991). We were somewhat disheartened by what appeared to be the inescapable conclusion that the only option for reaching a level of reproducibility in our ECG examination that would allow us to comfortably use the examination as a high stakes assessment in the clerkship was to substantially increase the number of cases. To help alleviate the burden of increasing the number of cases for both the students and the faculty, we experimented with using an extended matching format.

At about the same time, it dawned on us that a portion, perhaps a significant portion of the case-to-case variation in a student's performance might reflect differences in their ability to identify and manage different types of cardiac problems as apposed to manage different presentations of the same cardiac problem. These two sources of variation have different implications for the reproducibility of our examination. If a large part of what we were labeling case specificity reflected differences in the ability of a student to manage specific types of cardiac problems e.g. 1st degree AV block versus atrial fibrillation, rather than specific presentations of the same cardiac problem, the reproducibility of our examination might be higher than it appeared in our original set of studies. The nine cardiac problems and one normal ECG/patient scenario included in our original PBA largely covered the range of cardiac problems we expected our students to be able to identify and provide appropriate suggestions for initial management. In the language of generalizability theory, the variation in a student's performance among cardiac problems is a fixed facet and would not contribute to measurement error. Since there was only one case for each cardiac problem included in the original PBA, it was not possible to disentangle the variation associated with cardiac problems from the variation associated with different presentations of the same cardiac problem, hence the reason we conducted study 3.

The results of study 3 suggest the overwhelming source of variance in performance in our PBA is related to differences in performance among different presentations of a given cardiac problem rather than between different types cardiac problems. This suggests our original estimates of the generalizability of the exam were essentially correct. Along with being disappointing, this finding is puzzling and not at all what was expected. It seems intuitive that if students are able to identify the key features of a specific type of arrhythmia and have an

understanding of the appropriate management of the cardiac problem associated with the arrhythmia, there would have been more consistency in their performance across different presentations of the same cardiac problem. It should be noted the clerkship directors put a great deal of effort into locating ECGs for the exam that were classic presentations of a single, specific cardiac problems.

Limitations

This study has several limitations that should be noted in interpreting the findings. First, as often is the case in generalizability studies, the sample sizes in terms of cases, problems, and students were quite small, particularly in studies 1 and 2. We provided standard errors for the estimated variance components to help in the interpretation of the results however the standard errors themselves should be interpreted with some caution as they are not robust in terms of violations of the assumption of normality. At the same time, while the individual estimates of the variance components may reflect a fair amount of sampling error, the consistent pattern of case specificity being the overwhelming source of measurement error seems quite clear and probably real. Given the magnitude of the differences, it also seems likely that our finding that case specificity is largely due to differences in the performance of individual students across cases within types of cardiac problems rather than between cardiac problems is also probably real though it remains puzzling and we believe deserves further study. The extent these findings are applicable to other types of PBAs also remains an open question.

Lessons Learned

We feel this study provides several important lesions.

- When free response measures are used in PBAs, rating criteria need to reflect not only the scoring of right or partially right responses, but also how to address clearly wrong responses.
- Improving inter-rater reliability, at least for our particular PBA was of little value as the overwhelming source of error in the ratings was due to case specificity.
- It is important to carefully think through how to conceptualize the sources of variation in performance-based assessments. While as it turned out, the variation among the clinical problems and the interaction of students and clinical problems accounted for a very small portion of the variance in the scores, if it had turned out to be larger, we would have significantly underestimated the reproducibility of our PBA.
- Most importantly, the psychometric properties of PBAs, like all types of assessments must be carefully considered when using them for making high stakes decisions about trainees. We are doing our trainees, ourselves and the public a serious disservice if we make important decisions based on examinations that appear to have a great deal of relevance but lack reproducibility.

References

Brennan RL. Generalizability Theory. Assessment Systems Corporation, St. Paul Mn, 2001.

Crick JE, Brennan RL, A general purpose analysis of variance system. Version 2.2, 1984, American College Testing Service.

Downing S. (2000) Assessment of knowledge with written test forms. in Norman GR, van der Vleuten CPM Newble DI. (eds) *International Handbook of Research in Medical Education*. pp. 647-672. Dordrecht/Boston/London: Kluwer Academic Publishers.

Mavis BE, Henry RC, Ogle KS, Hoppe RB. (1996) The emperor's new clothes: The OSCE reassessed. *Academic Medicine* May;71(5):447-53.

Swanson D.B, Norcini J.J. (1989) Factors influencing reproducibility of tests using standardized patients. *Teaching and Learning in Medicine* 1: 158-66.

Swanson, D.B., Norman, G.R., Linn, R.L. (1995) Performance-Based Assessments: Lessons from the Health Professions. *Educational Researcher* 24(5): 5-35.

Van Thiel, J., Kraan ,H.F., Van Der Vleuten, C.P.M. (1991) Reliability and feasibility of measuring medical interviewing skills: the revised Maastricht history-taking and advice checklist. *Medical Education* 25: 224-229.

Table 1

Estimated Sources of Variance in ECG Rating Studies 1 and 2

	First Study		Second Study			
Source			**New Raters**		**Repeat Raters**	
Students	**0.78**	**±0.76***	**0.48**	**±0.50**	**0.43**	**±0.40**
Case (ECG)	0.61	±0.76	0.24	±0.44	0.72	±0.54
Rater	0.47	±0.39	0.21	±0.14	0.17	±0.12
SxC	2.40	±0.90	3.41	±0.83	2.23	±0.55
CxR	0.22	±0.23	0.00	±0.05	0.18	±0.08
SxR	0.00	±0.08	0.02	±0.04	0.10	±0.06
SxCxR	1.28	±0.29	1.07	±0.12	1.06	±0.11

*Standard Error

Table 2

Estimated Sources of Variance in ECG Rating Study 3

Students	**0.26**	**±0.91***
Cardiac Problem	0.00	±0.49
Cases:Problems	1.28	±0.70
SxP	0.03	+0.17
SxC:P	3.35	±0.22

*Standard Error

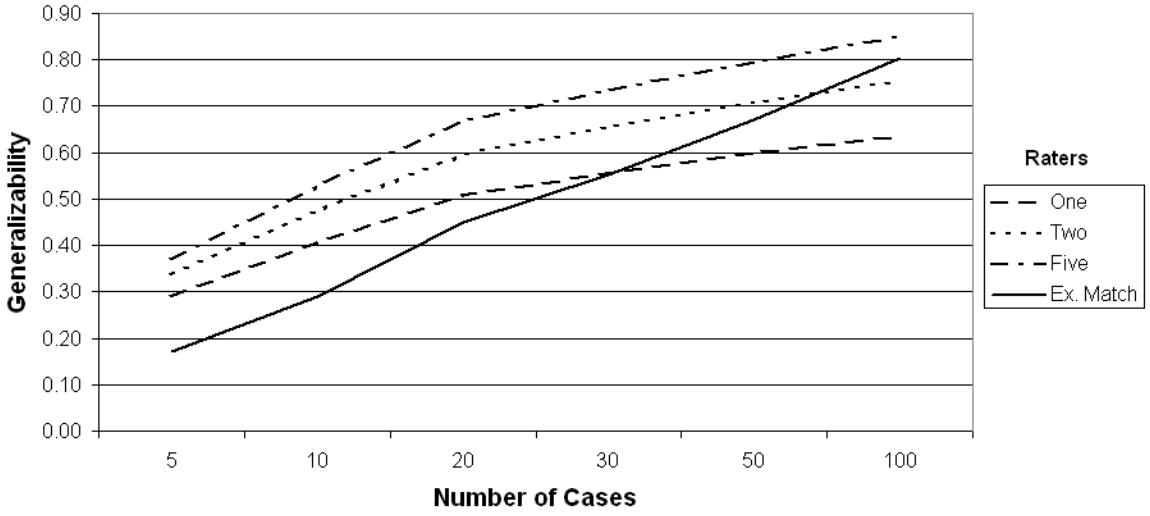

Figure 1 Caption

Generalizability for Different Numbers of Raters and Cases Using a Short-Answer (Study2) and Extended Matching (Study 3) Formats