Canadian Medical Education Journal

*Major Contribution/Research Article*

# Peer and Self-Assessment of Professionalism in Undergraduate Medical Students at the University of Calgary

*Pauline Alakija and Jocelyn Lockyer*

University of Calgary, Canada

Published: September 30, 2011

CMEJ 2011, 2(2):e65-e72 Available at http://www.cmej.ca



## Abstract

**Background:** Peer and self assessment processes are integral to the development of professional behaviours. The purpose of this study was to assess the Rochester Peer Assessment Tool (RPAT) among a group of volunteer first year students.

**Methods:** We assessed feasibility through participation rates. The evidence for the validity of instrument scores was ascertained through an exploratory factor analysis, MANOVA to determine age and gender differences, and a discrepancy analysis between the self and peer data. Reliability analyses included the Cronbach's alpha analysis and G- and D-studies. Students completed a feedback questionnaire to provide data about acceptability.

**Results:** Self and peer data were collected for 46 and 44 students, respectively. Each student had a mean of 7.2 peer assessments (out of a possible 8). The factor analysis identified two factors, interpersonal skills and work study habits. The discrepancy analysis showed students in the lowest/highest quartiles, as assessed by peers, had higher/lower self means than peer means. The G-coefficient was $Ep^2 = 0.77$. Student feedback was positive.

**Conclusions:** RPAT was feasible in our setting, was acceptable to the students, and has been adopted as a mandatory part of our program for first and second year students. The study added to the evidence base for the reliability and validity of the RPAT instrument scores as a method of assessing professional behaviours.

Correspondence: Dr. Jocelyn Lockyer Associate Dean, Continuing Medical Education and Professional Development, University of Calgary, 3280 Hospital Drive NW, Calgary, Alberta, Canada T2N 4Z6; E-mail: lockyer@ucalgary.ca.

## Introduction

The teaching and assessment of professional behaviours are recognized as integral components of undergraduate and post-graduate curricula. The General Medical Council competencies in the United Kingdom;[1] the Accreditation Council for Graduate Medical Education competencies in the United States;[2] and the Royal College of Physicians and Surgeons of Canada CanMEDs roles[3] all attest to the need to explicitly teach and assess professionalism. While definitions vary, the CanMEDS description of the professional role notes that it is guided by codes of ethics and a commitment to clinical competence, the embracing of appropriate attitudes and behaviours, integrity, altruism, personal well-being, and to the promotion of the public good within their domain. These commitments form the basis of a social contract between a physician and society.[3]

Measuring professional behaviours can be difficult. There are few opportunities to observe professional behaviours. As Stern[4] (p 6) noted, "most faculty are well aware that they observe students only on their best behaviour and that they have limited opportunities to see students in realistic settings". Multisource feedback, which draws on the perspectives of other physicians, other health care professionals, patients and a self-assessment, has been used to assess observable professional behaviours in postgraduate trainees and practicing physicians.[5] The professional behaviours of undergraduates, particularly in the pre-clinical years, are often only assessed by small group leaders and the occasional preceptor whom they shadow. Their peers, who are in a position to observe professional behaviours, are rarely asked to provide feedback. As a result, the opportunity to learn the skills of assessing others and receiving feedback from peers is delayed until the clerkship or post-graduate training commences. This delay may compromise acceptance of this type of data and partially explain why physicians may resist making changes based on this feedback.[6]

The University of Calgary Medical School curriculum is an innovative 3 year clinical presentation curriculum. In this structure, teaching in the first two years is organized around the 120 +/– ways that patients can present to a physician. These clinical presentations can take the form of a symptom (e.g., chest pain), physician examination signs (e.g., hypertension), or laboratory abnormalities (e.g., elevated serum lipids).[7] During the third and final year of medical school, the students complete their clerkship. The goals, educational objectives and operating philosophy of the school establish the expectation that students will be evaluated through a process that measures professional behaviour and includes peer assessment of the attainment of educational and professional objectives. Students are also expected to demonstrate self-directed lifelong learning skills, of which self-assessment is a fundamental component.[8]

Despite the goals outlined in the objectives and operating philosophy,[8] peer and self assessment skills were not being assessed in a consistent manner. Following a literature review and discussion within the Faculty, it was agreed that a study with volunteer medical students be conducted to determine whether the Rochester Peer Assessment Tool (RPAT) was an acceptable and feasible tool to assess professionalism in the pre-clinical program. The RPAT was specifically designed to provide peer feedback about professionalism.[9] Extensive testing in the Cleveland and Rochester settings was done by the instrument originators in their own schools.[9-14] These studies demonstrated that the RPAT had good psychometric properties. Furthermore, the instrument had been adapted into the 1st person so it could be used as a self-assessment tool.[11] An instrument with both peer and self components is useful with assisting students in calibrating their professional behaviours against the assessments provided by colleagues. Furthermore, this approach to receiving feedback from colleagues was in alignment with the feedback which practicing physicians receive from their colleagues. In Alberta, all physicians participate in a multi source feedback program every five years. The physicians receive feedback from colleagues, co-workers and patients and complete a self-assessment.[15,16] While the RPAT questionnaires have a different focus, RPAT has sufficient similarities to the instruments used by the College of Physicians and Surgeons of Alberta, Physician Achievement Review Program,[15,16] to enable students to learn the value of peer and self assessment at an early stage in their medical careers.

The purpose of the present study was to test the RPAT and a self-assessment version of the RPAT among a

group of volunteer 1st year medical students prior to including it within the university's assessment and feedback framework. We were interested in the feasibility, acceptability and the psychometric properties (i.e., evidence for validity and reliability) of the scores for both instruments in our setting. In the psychometric analysis, we extended the work done in Rochester and Cleveland [9-14] by examining the discrepancy between self and peer data. Work done previously in Alberta with practicing physicians showed that most physicians rated themselves as 'average' regardless of peer assessment ratings.[16] For those responsible for providing feedback and mentoring young clinicians, it can be helpful to know how students rate themselves relative to others when determining the feedback to provide.

## Method

### *The Instrument*

The RPAT is a 15-item survey (see Table 1 for list of items). The items are behaviourally anchored. Students were rated by their peer assessors on a 1-5 point scale with the option of an "unable to assess". Scores of 1 or 2 were low/ unsatisfactory. Scores of 4 or 5 were high/ exceptional.

### *Participants and Procedures*

All 152 students in the graduating class of 2010 were invited to participate in the study during the latter half of their first year. Students provided the names of 8 classmates. Students named their own assessors because previous experience with practicing physicians has shown that peer ratings are not biased substantially by the method of selection of the peers or the relationship between the rater and the subject.[17] This is the approach that has been adopted for use with practicing physicians as it appears the person being assessed is in the best position to identify the people who can observe them.

Both students and their peer assessors received e-mails with a link to the survey. Data were collected between February and July of 2008. Students received their feedback during an individual interview during the first half of the 2nd year with one of the authors (PA) who discussed the results of the survey with them and asked each student to complete a 7-item feedback questionnaire.

### *Assessment of the Instrument*

In order to fully understand the RPAT and its functioning in our setting, a number of assessments were conducted. Both feasibility and acceptability were examined. The scores were also assessed for evidence of validity and reliability.

First the feasibility of both instruments was assessed through participation rate, number of peer assessors per student, and "unable to assess" responses to items. The means and standard deviations were computed for all items.

Evidence of validity for the peer instrument scores were assessed using factor analysis. Prior work had established that the instrument had 2 factors (work habits and interpersonal attributes). For our study, an exploratory factor analysis (EFA) on the peer assessment data was completed as this would allow underlying dimensions or constructs to be extracted from a large set of variables; thus reducing multiple variables to a few factors or components. To determine whether there were differences in the ratings/scores related to sex and/or age, a multivariate analysis of variance (MANOVA) was calculated on the self assessment and peer assessment data. The MANOVA is a statistical technique used to assess group differences across multiple metric-dependent variables simultaneously, based on a set of non-metric factors acting as independent variables. Thus, a MANOVA calculation can determine if sex or age, or interactions between sex and age are related to the scores obtained in the peer assessments. This was considered important because assessment tools should be 'neutral' and not affected by student demographic characteristics. For the discrepancy analysis, we followed the approach used in a similar analysis of practicing physician data.[16] Each student was ranked 1 to 44 using the peer and assessment scores for each factor, and then the same was done for self scores. Each student was given a percentile ranking according to the raw peer and self scores. The class was then divided into 4 quartiles according to the percentile rank assigned to that student using the peer scores. Each quartile had a mean percentile score calculated for both the peer and self percentiles. The mean percentiles were

compared for both peer and self assessment scores for each quartile and each factor.

Reliability was initially examined by calculating the Cronbach's α for both instruments. A Generalizability study (G-study) was conducted by calculating the G-coefficient ($EP^2$) of the peer instrument to determine whether the number of items ($n$ = 15) and numbers of peer assessors ($n$ = 8) achieved sufficient data stability for an individual participant. In a G-study, a G-coefficient that is between 0 and 1 is calculated; this number represents the dependability of relative scores based upon student average per question scores. The G-analysis was based upon a single-facet nested design with raters nested within the students who were assessed using the following formula:[18]

$$G = E\rho^2 = \frac{Student\ (\text{var}\ comp)}{Student\ (\text{var}\ comp) + Error\ (\text{var}\ comp)}$$

A Decision study (D-study) was also conducted to determine the dependability of the scores if the number of raters was varied based upon 4 – 10 raters. D-studies are used to explore the impact of making design changes on the G-coefficient. The D-study was used to determine the number of peer assessors needed to obtain an acceptable G-coefficient. In formative feedback studies, G-coefficients in the range of 0.7 – 0.8 appear to be acceptable.[5] Having fewer assessors (i.e., 7, not 8 or more) required for dependability is desirable as this means fewer resources are required to produce dependable data. There is a greater likelihood of obtaining enough peer raters as some students may not choose to participate as peer raters.

Acceptability was assessed from the student feedback questionnaire by querying the utility of preparation sessions and written instructions; comfort with the anonymity and confidentiality procedures; respectfulness of the assessment; the ability of the feedback to motivate improvement, and willingness to participate in future peer assessments. These items were assessed on a 1-5 scale (strongly disagree to strongly agree). For these data, the mean and standard deviation were calculated for each item.

***Ethics***

The study was approved by the University of Calgary Conjoint Health Research Ethics Board.

## Results

All 152 students in the class of 2010 were given an opportunity to participate. Of these students, 146 were newly admitted in 2007. There were 64 male students (42%) and 88 female students (58%); the mean age was 24.8 years. Of the 152 students, 52 signed voluntary and informed consents to participate; 46 students completed the process of self evaluation. Both self evaluation and peer assessment data were collected for 44 students. Of these 44 students, 14 (32%) were male students and 30 were female students. There were 19 (43%) who were less than 25 years of age and 25 were 25 years or older. A total of 316 out of 368 (85.9%) of peer surveys were collected with a mean of 7.2 out of 8 peers assessing each student. Of these 44 students, 43 participated in the one-to-one interview with the investigator and 33 completed the student feedback questionnaire.

As shown in Table 1, the peer means were all > 4.5 out of 5. Self assessment mean scores ranged from 3.52 to 4.57. The EFA revealed a two factor solution, related to interpersonal skills and work study habits, which accounted for 64.68% of the variance. The varimax rotation converged in three iterations. The MANOVA, calculated to determine whether there was a difference on the self and peer assessments by sex and age did not reveal main or interaction effects for either the full scale or either of the two factors. Discrepancy analyses were conducted for each factor.

As shown in Table 2, students in the lowest quartile as assessed by peers had higher self rankings than peer rankings while students in the highest quartile had lower self rankings than peer rankings. Students, regardless of quartile, assessed themselves similarly, as about average.

The Cronbach's alphas for the self assessment and peer assessments were $\alpha$ = 0.82 and 0.92, respectively. The G-coefficient for the peer assessment with 8 assessors for a 15 item scale was $Ep^2$ = 0.77. The D-study produced G-coefficients of 0.62 (4 assessors), 0.67 (5 assessors), 0.71 (6 assessors), 0.74 (7 assessors), 0.77 (8 assessors), 0.79 (9 assessors), and 0.80 (10 assessors).

**Table 1. Descriptive Statistics and Factor Analysis for 316 Returned Peer Assessments for 44 Students and 46 Self-assessments.**

| RPAT Questions (score 1-5) | | Peer Data | | | | | Self Data | | |
|---|---|---|---|---|---|---|---|---|---|
| **Behaviours for low scores** | **Behaviours for high scores** | ***N*** | **Mean** | ***SD*** | **Factor 1** | **Factor 2** | ***N*** | **Mean** | ***SD*** |
| 1. Consistently seems unprepared for sessions; presents minimal amount of material; seldom supports statements with appropriate references | Consistently well prepared for sessions, presents extra material, supports statements with appropriate references | 290 | 4.47 | 0.60 | .30 | **.80** | 46 | 3.52 | 0.69 |
| 2. Overlooks important data and fails to identify or solve problems correctly | Identifies and solves problems using intelligent interpretation of data | 305 | 4.58 | 0.54 | .25 | **.81** | 46 | 3.98 | 0.58 |
| 3. Unable to explain clearly his or her reasoning process with regard to solving a problem, basic mechanisms, concepts etc. | Able to explain clearly his or her reasoning process with regard to solving a problem, basic mechanisms, concepts, etc. | 305 | 4.59 | 0.56 | .19 | **.81** | 46 | 4.00 | 0.70 |
| 4. Lacks appropriate respect, compassion and empathy | Always demonstrates respect, compassion and empathy | 315 | 4.69 | 0.56 | **.91** | .08 | 46 | 4.50 | 0.55 |
| 5. Displays insensitivity and lack of understanding for others' views. | Seeks to understand others' views | 310 | 4.58 | 0.63 | **.87** | .03 | 46 | 4.57 | 0.54 |
| 6. Lacks initiative or leadership qualities | Takes initiative and provides leadership | 305 | 4.50 | 0.67 | .13 | **.64** | 46 | 3.98 | 0.88 |
| 7. Doesn't share information or resources; impatient when others are slow to learn; hinders group process; tends to dominate group | Shares information or resources; truly helps others learn; contributes to the group process; able to defer to the group's needs | 299 | 4.63 | 0.61 | **.64** | .42 | 46 | 4.17 | 0.68 |
| 8. Only assumes responsibility when forced to or stimulated for personal reasons; fails to follow through consistently | Seeks appropriate responsibility; consistently identifies tasks and completes them efficiently and thoroughly | 301 | 4.62 | 0.57 | .14 | **.79** | 46 | 4.04 | 0.70 |
| 9. Does not seek feedback; defensive or fails to respond to feedback | Asks classmates and professors for feedback and then puts suggestions to good use | 264 | 4.45 | 0.65 | **.78** | .22 | 46 | 3.82 | 0.76 |
| 10. Pleases superiors while undermining peers; untrustworthy | Presents him/herself consistently to superiors and peers; trustworthy | 312 | 4.72 | 0.53 | **.74** | .45 | 46 | 4.50 | 0.55 |
| 11. Hides his or her own mistakes; deceptive | Admits and corrects his or her own mistakes, truthful | 303 | 4.71 | 0.55 | **.83** | .32 | 46 | 4.43 | 0.62 |
| 12. Dress and appearance often inappropriate for the situation | Dress and appearance always appropriate for the situation | 316 | 4.79 | 0.46 | **.42** | .32 | 46 | 4.52 | 0.59 |
| 13. Behaviour is frequently inappropriate | Behaviour is always appropriate | 315 | 4.74 | 0.48 | **.61** | .10 | 46 | 4.33 | 0.56 |
| 14. Dependent upon others for direction with regard to his or her learning agenda. | Directs own learning agenda; able to think and work independently | 298 | 4.70 | 0.55 | .18 | **.82** | 46 | 4.35 | 0.67 |
| 15. I have concerns for his or her future patients | I would refer my own family or patients to this future physician or ask this person to be my physician | 314 | 4.73 | 0.52 | **.79** | .32 | 46 | 4.40 | 0.67 |
| Cronbach's alpha | | | | | .79 | .78 | | | |
| % of variance accounted for | | | | | 49.43 | 15.25 | | | |
| Eigenvalues | | | | | 7.42 | 2.29 | | | |

**Table 2. Comparison of Means for Peer and Self Percentiles for Quartile Groups**

| Scale | Rater | Mean percentiles in each quartile | | | |
|---|---|---|---|---|---|
| | | 1st | 2nd | 3rd | 4th |
| Interpersonal skills | Peer | 13.85 | 38.64 | 63.64 | 88.84 |
| | Self | 47.31 | 59.91 | 52.06 | 62.20 |
| Work/study habits | Peer | 13.64 | 38.64 | 63.85 | 88.83 |
| | Self | 40.10 | 56.61 | 66.53 | 64.43 |

The students provided positive feedback about the assessment. All items from each student were assessed within a range of 3-5 out of 5 as shown in Table 3. The students were positive about the experience, appeared fairly motivated by the feedback to make changes in behaviour and expressed a willingness to participate in future peer assessment exercises.

**Table 3. Descriptive Statistics from the Student Feedback Survey**

| | *N* | Min | Max | Mean | *SD* |
|---|---|---|---|---|---|
| The preparation discussions helped me to fill out the peer and self assessment forms | 30 | 3 | 5 | 4.10 | 0.61 |
| The written instructions helped me fill out the peer and self assessment forms | 32 | 4 | 5 | 4.28 | 0.46 |
| The process preserved my anonymity (my colleagues and preceptors did not know how I felt in the questionnaires). | 33 | 3 | 5 | 4.79 | 0.49 |
| The process preserved by confidentiality (my results were not known by my colleagues or the Undergraduate Medical Education Office) | 33 | 4 | 5 | 4.85 | 0.36 |
| The process was respectful to me | 33 | 4 | 5 | 4.82 | 0.39 |
| The feedback I received has motivated me to make positive changes in how I behave | 31 | 3 | 5 | 3.94 | 0.63 |
| I would be willing to participate in future peer assessment exercises | 33 | 3 | 5 | 4.58 | 0.61 |

## Discussion

This study showed that the RPAT was a feasible and acceptable method of evaluating professional behaviours and initiating a formal structured process for doing peer and self assessments. We continued to build the evidence for the validity and reliability of the instrument scores for both peer and self versions of RPAT in a 3-year medical school. We extended the psychometric assessment with a discrepancy analysis to examine how students in different quartiles as measured by their peers, assessed themselves.

The assessment was feasible. Students stated that they were prepared to handle the self and peer assessments. Response rates were acceptable as our goal was 50 participants; this was an initial study and an extensive de-brief with students was an integral component of the work required to establish the viability of this type of assessment for the school. Furthermore, the G-coefficient ($Ep^2$ = 0.77) with 8 assessors shows that this was feasible administratively and acceptable from the perspective of a formative assessment. Increasing the number of assessors to 10 would produce a G-coefficient >.80. However, these data suggest that a minimum of 6 assessors would produce an acceptable reliability ($Ep^2 \geq 0.70$).[5]

RPAT was also acceptable to the students who stated it was respectful and preserved their confidentiality and anonymity. The students indicated they would be willing to participate again in an RPAT assessment in the future.

We have added to the evidence for the validity of the RPAT instrument scores. We re-confirmed the factors identified previously.[9] The reliability analyses for the two factors indicated that the Cronbach's alpha calculations were close to $\alpha$ = .80. The MANOVA indicated that scores do not differ according to age and sex characteristics. By adding a self assessment instrument, students had additional data they could use to consider how well they were doing. For those mentoring students, having both self and peer data can be helpful in assisting the students calibrate their self-assessment. These data allowed us to conduct a discrepancy analysis of self and peer data. Similar to the findings from an analysis done with practicing physicians,[16] this analysis showed that most participants evaluated themselves as 'average' and that those in the first and fourth quartiles provided

the least accurate self-assessments. Not all students stated that they were motivated by the assessment. The items on the student feedback survey that had the lowest scores and the greatest standard deviation queried whether they were motivated to make positive changes in behaviour. It is possible that students who did well on this assessment, as judged by their peers, did not feel they could use the data.

There are limitations to this study. We limited the number of participants to 50 students, approximately 1/3 of the class. Student participation was voluntary. It is possible that those students who might have benefitted most from this type of assessment (i.e., those who perceived that their behaviours would be judged poorly, had previous conflicts with peers, had less well developed networks), may not have volunteered to participate. Although we surveyed students about their perceptions, we did not collect data from the individual de-briefings with the students. That decision was deliberate as we wanted to encourage participation and trust between the investigator and the students.

We believe the RPAT meets our criteria for assessment. It is feasible with 6 assessors. It was acceptable to those students who participated. There is evidence that the assessment is valid and reliable. RPAT has been adopted as a mandatory part of our assessment system for both first and second year students.

***Main Findings***

- Peer assessments give valuable feedback to undergraduate medical students about their professional behaviours
- There are significant discrepancies between self and peer evaluations regarding professional behaviours
- The RPAT peer assessment tool and adapted self assessment tool are feasible in a Canadian undergraduate medical school environment
- There is evidence for the validity and reliability of RPAT peer assessments in a Canadian undergraduate medical school

***Acknowledgements***

The authors thank R. Epstein, Rochester School of Medicine and Dentistry, for permission to use the RPAT; Dr Bruce Wright, Associate Dean, Undergraduate Medical Education, The University of Calgary for enabling the study to be done; members of Dr. Alakija's MSc Committee, Drs Keith Brownell and Gwen Hollaar, for their contributions and feedback during the undertaking of the study; and Tom Durnin for facilitating the programming required to conduct the study.

## References

1. General Medical Council. *Good Medical Practice*, November 13, 2006. http://www.gmc-uk.org/static/documents/content/GMP_0910.pdf [Accessed April 15, 2011].
2. Accreditation Council for Graduate Medical Education, *Common Program Requirements: General Competencies*. Approved by the ACGME Board, Feb 13, 2007. http://www.acgme.org/outcome/comp/GeneralCompetenciesStandards21307.pdf [Accessed April 15, 2011].
3. Royal College of Physicians and Surgeons of Canada, *The CanMEDS 2005 Physician Competency Framework*, Oct 25, 2010. http://www.rcpsc.medical.org/canmeds/ [Accessed April 15, 2011].
4. Stern DT, A framework for measuring professionalism, in DT Stern (ed), *Measuring medical professionalism*. Oxford: Oxford University Press, 2006.
5. Lockyer J, Clyman S. Multi Source Feedback, in E Holmboe & R Hawkins (eds), *A practical guide to the assessment of clinical competence*. Mosby/Elsevier, 2008.
6. Sargeant J, Mann K, Sinclair D, van der Vleuten C, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv in Health Sci Educ.* 2006;10.1007/s10459-006-9039-x.
7. Lockyer JM, Violato C, Wright BJ, Fidler HM. An analysis of long-term outcomes of the impact of curriculum: A comparison of the three- and four-year medical school curricula. *Acad Med.* 2009;84(10):1342-1347.
8. University of Calgary, Faculty of Medicine, *Operating Philosophy*, http://www.ucalgary.ca/mdprogram/operatingphilosophy [Accessed April 15, 2011].
9. Dannefer EF, Henson LC, Bierer SB, Grady-Weliky TA, Meldrum S, Nofziger AC, Barclay C, Epstein RM. Peer assessment of professional competence. *Med Educ.* 2005;39:713-722.

10. Lurie SJ, Lambert DR Nofziger AC, Epstein RM, Grady-Weliky TA. Relationship between peer assessment during medical school, Dean's letter rankings, and ratings by internship directors. *J Gen Int Med.* 2007;22:13-16.

11. Lurie SJ, Meldrum S, Nofziger AC, Sillin LF 3rd, Mooney CJ, Epstein RM. Changes in self-perceived abilities among male and female medical students after the first year of clinical training. *Med Teach.* 2007;29:921-926.

12. Lurie SJ, Nofziger AC, Meldrum S, Mooney C, Epstein RM. Effects of rater selection on peer assessment among medical students. *Med Educ.* 2006;40:1088-1097.

13. Lurie SJ, Nofziger AC, Meldrum S, Mooney C, Epstein RM. Temporal and group-related trends in peer assessment amongst medical students. *Med Educ.* 2006;40:840-847.

14. Nofziger AC, Naumburg EH, Davis BJ, Mooney CJ, Epstein RM. Impact of peer assessment on the professional development of medical students: A qualitative study. *Acad Med.* 2010;85(1):140-147.

15. College of Physicians and Surgeons of Alberta, Physician Achievement Review Program, www.par-program.org [Accessed April 15, 2011].

16. Violato C, Lockyer J. Self and peer assessment of pediatricians, psychiatrists and medicine specialists: Implications for self directed learning. *Adv Heal Sci Educ.*2006;11:235-244.

17. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA* 1993;269(13):1655-1660.

18. Brennan RL. *Generalizability Theory*. New York: Springer-Verlag, 2001