



# Best Practices in Teacher Assessment

Summary of Recommendations

2010

**Prepared by the Best Practices in Teacher Assessment Working Group (BPTAWG):**

Glen Bandiera (Chair), Kenneth Fung, Karl Iglar, Markku Nousiainen, Katina Tzanetos, Aikta Verma, Veronica Wadey  
Susan Glover-Takahashi, Erika Abner, Caroline Abrahams, Yaw Otchere, Mariela Ruetalo

## TABLE OF CONTENTS

Background .....	1
Methodology .....	1
Results.....	2
Analysis and Discussion .....	4
Recommendations.....	5
1. The Process of Teacher Evaluations .....	5
2. Form Content and Design.....	6
3. Next Steps .....	7
4. Other Recommendations .....	8
APPENDIX ONE: Terms of Reference and Work Plan.....	9
APPENDIX TWO – Elements of Good Teaching for Inclusion on All Forms.....	11
APPENDIX THREE: Potential Comments for Drop-Down Lists.....	12
APPENDIX FOUR – Sample Forms .....	14
APPENDIX FIVE: POWER Survey Summary Results.....	16
APPENDIX SIX: Relationships between TES and ITERS .....	21
APPENDIX SEVEN: References .....	31

# BEST PRACTICES IN TEACHER ASSESSMENT

## Summary of Recommendations

### BACKGROUND

The Royal College of Physicians and Surgeons of Canada (RCPSC) and the College of Family Physicians of Canada (CFPC) mandate the evaluation of teaching through accreditation standards. A valid and reliable assessment system for teaching is foundational to delivering the best educational experience to residents, informing decisions about teaching contributions of individual faculty members, and identifying faculty members who might benefit from assistance with teaching techniques.

The University of Toronto Postgraduate Medical Education programs have been collecting data on teacher evaluations using the POWER (the Postgraduate Web Evaluation and Registration) system since 2003. In 2010, the majority of teacher evaluations in PGME are undertaken through POWER yet the forms and processes used to collect, disseminate and act on the data varies considerably across programs and hospitals.

The POWER Steering Committee has reviewed the content of teacher evaluation forms across clinical departments and determined that there is great variability in the forms, the scores and completion rates. Teacher evaluation forms are currently designed without the benefit of identified PGME best practices. While a number of studies have identified good teaching practices in various settings, no compilation of qualities of a “good teacher” has been established to assist in institutional teaching assessment form design, process improvement, or comparative analysis. .

The design and implementation of best practices through a set of minimum standards would facilitate improved assessment of teaching and interpretation of results. Improved evaluation instruments could also increase teacher evaluation completion rates and allow for interdepartmental and intradepartmental comparisons. Such an institutional approach to application of best practices in teaching assessment has to our knowledge, not been done.

In 2009, the POWER Steering Committee recommended that a small working group be formed with the goal to develop a set of best practices in teacher assessment to inform guidelines that programs must use to develop their teacher assessment instruments and processes. The Best Practices in Teacher Assessment Working Group (BPTAWG) was formed. The committee, chaired by Dr. Glen Bandiera, a member of the POWER steering committee, also consisted of members of the University of Toronto community who were recommended for their knowledge or interest in teacher assessment (See Appendix A – Committee List). Practitioners from Family Medicine and several specialties and in different stages of their career, including Residents, were selected to represent a wide perspective.

### METHODOLOGY

The Work Plan adopted by the Working Group included a:

1. Review of the literature to examine the:
  - o process of teacher evaluations, and

- design of teacher evaluation forms, including evidence-based content areas for assessment.
2. Document review of:
    - the 2009 POWER User Satisfaction Survey results, and
    - sample forms recently deployed in the U of T POWER system
  3. Quantitative study to investigate the:
    - correlation between teaching assessment form length and completion rates at U of T, and
    - the correlation between ITER scores and TE scores
  4. Qualitative study of the comments provided on teacher evaluation scores

## RESULTS

The literature review yielded some important observations and analysis. *A Review of the Evaluation of Clinical Teaching: New Perspectives and Challenges* discusses aspects of the process of evaluation, methods and design and their importance for both teachers and the program. The review recommends adhering to basic measurement principles such as aiming for high levels of validity and reliability. To allow for comparisons, evaluations should be applicable to all levels and types of teachers, programs and sites. In addition, evaluation goals should be clearly defined and presented. A second recommendation is to include several perspectives. In addition to attaining feedback from learners, others' point of views should also be solicited such as that of teachers, patients and administrators. When asking learners for their feedback, it's important to only ask questions they are in a position to answer. Including a range of methodologies such as focus groups and interviews would also give the data fuller meaning. Lastly, the review recommends that evaluations should include attributes related to all the roles of a physician including CanMEDS.

A second review, *How Reliable Are Assessments of Clinical Teaching?*, aims to identify themes that may help in the development of meaningful assessment tools. Again, validity and reliability were discussed and various forms of each were described in detail. When considering categories of validity it should be understood that validity evidence exists to various degrees but there is no threshold at which an assessment is valid. The most frequent measure of reliability found in the literature was internal consistency of reliability of teaching domains. Altogether 14 domains of teaching were identified in the literature with the most common being clinical-teaching and interpersonal skills. Themes identified in developing meaningful assessment tools are outcome measures, the environment where the learning and evaluating exists and objectivity of the assessment of the faculty. Limitations identified include inflated ratings when using learners to evaluate faculty, which prompt the need for closer attention to comments written on evaluations; and the unique cultures of teaching at many institutions may limit the generalizability of even the most carefully designed evaluations.

*Evaluating the Performance of Inpatient Attending Physicians: A New Instrument for Today's Teaching Hospitals* discusses a thorough process of developing a new evaluation tool which would replace evaluations that were deemed unable to capture the diversity of clinical

teachers' responsibilities. Once testing reliability (by measuring the consistency among responses to items within each domain and the consistency among different residents when they evaluated the same physician on the same domain), and assessing the instrument's face, content and construct reliability, the authors developed a tool that measures 9 domains with several questions pertaining to each and a summary score which is the mean score of all questions. Residents and physicians appeared to be satisfied with the tool. The next steps are to test whether the findings are generalizable to other institutions. If so, the use of this instrument would allow credible evaluation of clinical faculty and to help standardize criteria for academic promotion.

*Are Anonymous Evaluations a Better Assessment of Faculty Teaching Performance? A Comparative Analysis of Open and Anonymous Evaluation Process* discusses the issue of anonymity in teacher evaluations; a topic not discussed in the review articles. The authors conducted research in which residents and medical students evaluated faculty, first using an evaluation in which their names were indicated and then completing an anonymous evaluation on the same faculty. They found a statistically significant difference between the anonymous and open evaluations. Faculty received lower scores on the anonymous evaluations across all items.

In late 2008 and early 2009, the Postgraduate Medical Education (PGME) Office at the University of Toronto conducted a survey on several uses of POWER focusing on functions related to evaluation. A number of recommendations arose from the teachers. Some highlights are:

- Comments should be required for very high or very low scores
- The forms should be shortened and the domains of evaluation should be condensed
- Behavioral questions should be used that do not ascribe motivation to actions
- Holding forms back to protect trainees creates a long feedback loop for teachers, too long to be useful
- Trainees and teachers should be blinded from the results of each other's evaluation until both are filled out.

Key themes that arose from the trainees' responses in the survey are:

- Forms are generally too long and the content is sometimes irrelevant
- The opportunity to include more qualitative comments is desirable
- There's a need for clarity around exactly how confidential the TES reports are
- Technical issues around the ease of filling out multiple forms and improving the accuracy of teacher lists

A quantitative study to investigate the relationship between the scores given to residents on In Training Evaluation Reports (ITERs) and the scores they gave their teachers on teacher evaluations (TEs) was conducted by the PGME Office for the BPTAWG. To answer the question, "Do trainees who receive lower ITER scores give lower teacher evaluations?" evaluations for internal rotations only were used and they were paired with the teacher evaluations from the same rotation. A positive correlation between ITER and teacher

evaluation scores were found (Wilcoxon Signed Rank,  $p=0.000$ ). For each point an ITER drops, TES scores decrease by 0.16 points. A look at whether the requirement to complete a teacher evaluation prior to receiving their ITER affects teacher evaluation scores found that programs which do not hide ITER results from a trainee until at least one teacher evaluation and one rotation evaluation have been completed, tend to have slightly higher average TES scores (0.11 points; Mann-Whitney,  $p=0.000$ ). Finally, an analysis to answer the question, “Do residents who receive detailed feedback about their rotation performance give higher or lower scores on their TE?” revealed that trainees who responded ‘yes’ to having detailed feedback rated their teachers higher than those who indicated that they do not receive feedback. Receiving feedback added an average of 0.07 points to a teacher evaluation (Mann-Whitney,  $p=0.000$ ).

## ANALYSIS AND DISCUSSION

The BPTAWG focused discussions on the processes involved in teacher evaluations and the content and design of teaching evaluation forms.

A correlation between low ITER scores and low TE scores exists regardless of the timing of the provision of relevant forms. (See Appendix Six). Although the reasons for this correlation may be multi-factorial, a perception persists among teachers that providing a resident with a critical ITER, even if accurate and justified, will place the teacher at risk of receiving a low TE score based inappropriately on the low ITER score rather than on effectiveness as a teacher (i.e., ‘retaliation’). The reliance on TE scores to determine promotions and remuneration, therefore, puts teachers who have this perception in a conflict when they are asked to accurately rate resident performance.

Other potential contributors to the correlation between low TE and ITER scores include the presence of a mutually negative teacher-learner relationship, and the perception that a low ITER score implies poor teaching and thus justifies a low TES score. Moreover, there is evidence that raters will base their assessments on an overall impression of a teacher, driven by factors that may or may not be explicitly asked about on the form. Therefore, final or overall scores may not reflect scores on individual items and if the individual items are too numerous, it is likely that the ratings for them will not be independent.

Residents should be clearly informed of the difference between anonymity and confidentiality. It is critical that resident confidentiality be maintained with respect to information provided to a teacher. However, resident anonymity should not exist. Programs have a responsibility to ensure that the teacher assessment process is functioning as intended, that teachers receive feedback in a regular, timely (at least annually) and formative manner, and that those who provide information on assessment forms can be held accountable for factual information provided. The provision of information on the forms can have significant impact on teachers and should be regarded as a professional obligation of residents. Teachers need to be protected from unfounded, unprofessional, irrelevant and/or egregious comments and ratings that are not based on experience or fact. Residents must be accountable for the assessments they provide, especially in those cases where the assessment calls the execution of the obligation to professionally assess a teacher is in question.

Teaching assessment forms are intended to enable Residents to provide their candid, complete and accurate perceptions and recollections of a teacher’s behaviour and impact.

While discrete teaching behaviours are the mainstay of effective teaching, it is recognized that the ability of teachers to role model professional practice beyond that of a teacher is a powerful way to influence learning. Residents should not only be taught explicitly what is to be expected of them as physicians, but should also be shown in a manner that reflects the current understanding of the role(s) of a physician.

Taking into account discussions held at BPTAWG meetings, the literature on teacher evaluations and results from the quantitative analyses, the BPTAWG has put forward recommendations for both the process and form content and design of teacher evaluations.

## **RECOMMENDATIONS**

### **1. The Process of Teacher Evaluations**

Teaching effectiveness scores can have a significant impact on remuneration, eligibility for teaching awards, promotions decisions, and demonstration of accountability to practice plans. The role of resident assessment of teaching is an important part of assessing the overall quality of teaching but its limitations must be taken into consideration when doing so.

- 1.1 Recognizing that the information gathered on teacher assessment forms is resident opinion and may or may not represent true effectiveness of one's performance as a teacher, the term 'Teaching Effectiveness Score' should not be used in reference to these forms in isolation. A more appropriate term, such as 'Resident Assessment of Performance as a Teacher (RAPT)' is preferred. The RAPT should be one of multiple means for assessing overall teaching effectiveness.
- 1.2 The PGMEAC should endorse a statement reflecting what constitutes effective teaching, especially as reflected in the items on the teaching assessment form.
- 1.3 There should be a well-publicized, easily understood and reliable appeals mechanism for teachers wishing to request an investigation into a teacher assessment score. Processes for including teacher assessments as a means to justify promotions or remuneration should provide due consideration for the removal of outliers. Departments and/or the faculty should identify expert assistance in establishing outlier criteria. Decisions on promotion and remuneration should not be made until final assessment data is reviewed.
- 1.4 Departments should establish a clear and well-publicized description of the process for how resident ratings of teachers will be used and what the implications of these ratings are. (e.g., at the time of initial faculty recruitment, engaging in a supervisory relationship, promotion, resident entry to programs).
- 1.5 Processes should include ongoing monitoring of resident assessments of teachers to ensure that urgent issues (potentially indicated by low scores) are identified quickly and also that comments on teacher assessment scores are professional and remain in the spirit of constructive feedback (monitored by program directors). This monitoring should



be done in such a way as to ensure confidentiality of resident identity to the teacher involved.

- 1.6 Departments should be responsible for identifying resources for teachers in need of professional development and should be supported in this by selected initiatives from the Faculty of Medicine. Resources should include basic faculty development programs and individualized teacher-specific interventions as deemed appropriate.

## **2. Form Content and Design**

The forms used to collect residents' perception of teaching should adhere to evidence-informed or best available practices. The forms should serve the dual function of allowing residents to provide their candid thoughts as well as highlight important aspects of good teaching. It is recognized that teachers often teach residents from various programs in a single environment and the ability to collate data from multiple residents is important. The ability for leaders, managers and administrators to collect and compare data in order to detect patterns, establish benchmarks, and recognize trends is an important function of a well-designed teacher assessment process.

- 2.1 The POWER Steering Committee should mandate the adoption of minimum standards for form design and content based on this report.
- 2.2 Forms used to collect teaching assessment data should:
  - Be of reasonable length
  - Define universally accepted good teaching practices
  - Allow some flexibility to incorporate program and environment-specific design elements
- 2.3 Residents should only be asked to comment on those elements of a teacher's performance that the Resident would reasonably be expected to be able to form an opinion on.
- 2.4 Forms should ideally include:
  - Instructions about proper form completion and a summary of implications of the data collected either on the top of each form or on a preliminary screen.
  - Two sections:
    - A generic portion that includes specific questions about pervasive teaching behaviours (See appendix Two)
    - Another portion specific to the specialty, the context, or both.
  - 7 to 8 quantitative questions (a maximum of 10)



- Of the 10 items, there must be a question related to overall teaching performance and the scale must clearly indicate a threshold for minimally acceptable performance. The most reliable form of Resident perception remains the overall cumulative impression the Resident has about the teacher (rather than specific scores on specific items).
  - Rating scales should provide five options, with feasible extremes and a neutral centre. The scales should indicate progressively positive ratings from left to right.
  - There should be a mandatory expectation that Residents provide justification for ratings of 1,2 or 5 on the five-point scale to encourage the provision of meaningful feedback, emphasize the importance of experience-based rating practices, and ensure due consideration of extreme ratings,. (May only apply to 5 core attributes)
  - The opportunity for Residents to add comments not reflected in the questions asked and to provide rationale for their choices of extreme ratings.
  - Some means by which Residents can comment on their teacher's performance as a role model. The role modeling function should be based on the CanMEDS paradigm.
- 2.5 Consideration should be given to providing drop-down menus of comments related to effective teaching behaviours to help Residents provide helpful feedback and to help teachers identify common themes in their assessments (See appendix Three).
- 2.6 Programs sending Residents to existing rotations that are administered and overseen by other departments/divisions should set these up as an external rotation to maximize the likelihood that the teacher list is up to date and to allow for the teaching assessment data to be collected on the same forms regardless of which program the Resident belongs to.

### **3. Next Steps**

The POWER Steering Committee is responsible for making recommendations about, and providing oversight of, the POWER system. The information in this report can be used to inform recommendations to the Vice-Dean PGME about modifications to the teacher assessment system. It is recommended that the POWER Steering Committee provide further commentary on this report and provide guidance regarding implementation of these recommendations and implied choices (for example, which of the elements in Appendix One to include in the teaching forms).. Practices vary across departments and programs. There is much uncertainty and anxiety about the process for collecting data on teachers and how the data is used. Broad support for a process and clear understanding of the rationale and design elements are critical to widespread acceptance and optimal use of the system.

- 3.1 The POWER Steering Committee should use formal links with the MedSIS implementation committee to ensure compatible and, where possible, synergistic strategies for adoption and implementation of recommendations.
- 3.2 To facilitate adoption, consideration should be given to soliciting input and support from a breadth of stakeholders, including POWER Steering Committee, PGMEAC (and constituents), HUEC (and constituents), and all-chairs.
- 3.3 Programs and departments should identify an individual responsible for supporting development of teacher assessment protocols and forms across programs in the department.
- 3.4 The PGME office should identify an individual or group who will review and provide commentary on the forms and processes.
- 3.5 The POWER Steering Committee should assume responsibility for monitoring compliance on form and process design and report this regularly to program directors and departmental leadership.
- 3.6 The implementation of the recommendations should proceed over an academic cycle (i.e., - one year) to enable adequate time for design, discussion and dissemination of new changes and to allow the deployment of new tools and processes to coincide with the start of an academic year.

#### **4. Other Recommendations**

- 4.1 The use of teacher effectiveness scores to determine remuneration and teaching awards in isolation should be discouraged. (include peer review, innovation, awards, quantity, breadth, etc) A more comprehensive method that includes, but is not overly reliant on, resident perception is preferred.

## **APPENDIX ONE: TERMS OF REFERENCE AND WORK PLAN**

### Best Practices in Teacher Assessment (BPTA) Working Group

#### Mandate

To provide advice to the POWER Steering Committee and the Vice Dean Postgraduate Medical Education about Best Practices in Teacher Assessment for postgraduate medicine at University of Toronto

#### Reporting to:

POWER Steering Committee through BPTA Working Group Chair

#### Responsibilities

The Working Group will:

- Advise the POWER Steering Committee on the best practices for evaluation of teacher effectiveness in such areas as:
  - Development of minimum design standards, including minimum best teaching practices, to assess the effectiveness of clinical teachers in PGME by June 2010,
  - Implementation of standardized TES forms for use within the POWER system for 2011-12 academic year.
- Review draft documents and provide feedback
- Recommend Implementation strategies

#### Composition

The BCTES Workgroup will include:

- Chair (Dr. Glen Bandiera)
- 6-7 members (i.e. including up to additional 2 POWER Steering Committee members)
- PGME Staff (e.g., Caroline, Sue GT, Erika, Mariela, Yaw)

#### Term of Office

The term of office for the BPTES Workgroup is for the duration of the BPTES project. The expected completion date of the project is fall 2010.

#### Meeting Format

- Face to face and/or teleconference meetings will be held 4-6 times as needed;
- Communication via email between teleconferences.
- A consensus model is used to arrive at the preferred course of action.

#### Confidentiality of Information

Workgroup members shall not divulge information that is revealed to them through work on this project, including communication/consultation with external organizations.

***Approach to completing objectives:***

1. Analyze characteristics and behaviors of effective clinical teachers.
  - a. Review literature on effective clinical teaching, having regard to different skills required in different medical settings.
  - b. Review, code and analyze qualitative comments provided by residents as part of the Teaching Effectiveness components of POWER to determine characteristics and behaviours associated with effective teaching:
    - i. Across different programs
    - ii. Across different types of teaching environments, which might include, for example, OR, Emerg, Ambulatory, Wards.
  - c. Develop a core list of effective characteristics and behaviours as well as program/environment-specific lists
2. Develop template teacher evaluation surveys to be used by all programs, based in part on research and analysis from #1. The surveys must meet agreed to standards, such as:
  - a. Provide useful feedback to teachers
  - b. Allow identification of problems
  - c. Allow identification of “star” teachers
  - d. Provide data that can be standardized across teaching sites and over time.
3. The teacher evaluation surveys will be crafted into two parts: a standard part for all programs and a department-program specific part that accurately reflects the teaching environment. The surveys will include “pull down” menus of qualitative phrases that residents can choose to describe teachers, in addition to open text boxes.

## APPENDIX TWO – Elements of Good Teaching for Inclusion on All Forms

Specific questions reflecting important teaching behaviours should form the basis for the assessment form. These should be rated on a five point scale, with least desirable performance on the left and best on the right. Providing some descriptors of appropriate behaviour for each rating level is desirable.

### ***This teacher:***

- 1. Made him/herself available to me so I had the support I needed.***
- 2. Ensured we agreed on expectations early on and did their best to meet them***
- 3. Encouraged me to explore my limits safely***
- 4. Provided regular, meaningful, prompt feedback to me***
- 5. Demonstrated respect for me as a learner and as a person***
- 6. Had the following overall impact on me as a learner***

In addition to the above, it is recommended that programs customize forms to the individual teaching environment. This may involve the design of forms specific to the following distinct teaching contexts:

- 1. Ward-based teaching***
- 2. Operating/procedure room teaching***
- 3. Outpatient Clinic Teaching***
- 4. Emergency Department Teaching***
- 5. One-on-one teaching (research/project/mentorship)***
- 6. Formal teaching (Workshops/seminars/lectures)***

In addition to the above, it is recommended that programs ask residents to comment on the physicians' impact as a role model on the CanMEDS competencies: These can either be open narratives or selections from a drop-down list. (See Appendix Two).

Medical Expert  
Collaborator  
Communicator  
Health Advocate  
Manager  
Scholar  
Professional

## APPENDIX THREE: Potential Comments for Drop-Down Lists

To assist residents in providing constructive assessments and to minimize the time required to complete the form, it is recommended that a series of drop down lists be created to allow for comments on the Role Model aspects of being a good teacher. Each item can have a paired drop-down item with the negative version – ‘did not’. Each list should be preceded with the statement: “This teacher...” (Maybe phrased as “This teacher showed me how to...” or “This teacher made me want to adopt the following things they do...” (With appropriate re-phrasing of the items below).

### **Medical Expert**

*Demonstrated critical thinking*  
*Demonstrated a breadth of knowledge about their specialty area*  
*Demonstrated excellent physical examination skills*  
*Modeled effective diagnostic reasoning*  
*Was adept at procedures*  
*Was able to acknowledge their limitations openly*  
*Provided humane patient care*  
*Demonstrated excellent clinical judgment*

### **Collaborator**

*Respected and maximized the roles of other team members*  
*Addressed conflict effectively*  
*Judiciously involved consultants or engaged in consultations appropriately, respecting the role of other physicians in patient care*  
*Delegated responsibilities clearly*  
*Respected diversity*  
*Committed to team learning*  
*Handled a team well in times of crisis*  
*Embraced shared decision-making*  
*Involved other agencies effectively in patient care*

### **Communicator**

*Provided concise information*  
*Took steps to ensure that accurate information was received*  
*Demonstrated effective communication in the face of language, cultural or hearing impediments*  
*Took steps to ensure all team members were informed of plans and decisions*  
*Ensured that I was informed of upcoming events and expectations*  
*Used effective written communication*  
*Spoke to me and others with respect*  
*Demonstrated patience and perseverance during difficult communications*

*Used electronic media responsibly and effectively*  
*Able to establish an effective physician-patient relationship*  
*Was a good listener*  
*Adept at disclosing error*  
*Adept at breaking bad news*

### **Health Advocate**

*Demonstrated concern for overall patient wellbeing*  
*Took steps to address impediments to good patient care*  
*Demonstrated a focus on risk factors and preventive medicine*  
*Was vigilant for medical error and took an open, constructive approach when errors occurred*  
*Demonstrated knowledge and concern for specific health issues affecting practice population*  
*Was focused on patient safety*  
*Demonstrated awareness of how policy affects care*  
*Was able to use their influence and power to achieve important goals*

### **Manager**

*Was respectful of team members' time*  
*Adhered to time commitments*  
*Used resources and time well*  
*Encouraged clear expectations, role awareness and accountability in teams.*  
*Involved me in quality improvement*  
*Assumed a leadership role appropriately*  
*Ran effective meetings*  
*Was able to rally support for a cause*

### **Professional**

*Demonstrated self control*  
*Demonstrated ethical practice and integrity*  
*Demonstrated attention to personal wellbeing and balance*  
*Demonstrated sensitivity to gender, ethical, cultural, and socioeconomic issues*  
*Acted with honesty and integrity*  
*Acted in the patient's best interests*  
*Held high standards of practice*  
*Demonstrated respect for professional regulatory authorities and institutional regulations*

### **Scholar**

*Consistently used relevant evidence to inform practice*  
*Was able to identify relevant issues that arose in the literature around topics*  
*Did a good job of using all available information to solve problems*  
*Embraced discussions of alternate points of view*  
*Demonstrated a commitment to lifelong learning*





**Part C:** Role modeling is a critical source of learning. Recognizing this, please select from the drop-down lists up to three items that you want to adopt from this teacher for your practice and up to three items that you feel this teacher should consider improving for EACH of the CanMEDS Roles listed below

CanMEDS Role	I will adopt	I will adopt	I will adopt	Can Improve	Can Improve	Can Improve
Medical Expert	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list
Communicator	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list
Collaborator	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list
Health Advocate	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list
Manager	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list
Professional	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list
Scholar	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list	Drop down list
Other comments	Open Narrative					

## APPENDIX FIVE: POWER Survey Summary Results

### 2008/09 POWER User Survey: Teacher Evaluation Responses by Teachers Policy and Analysis – February 2010

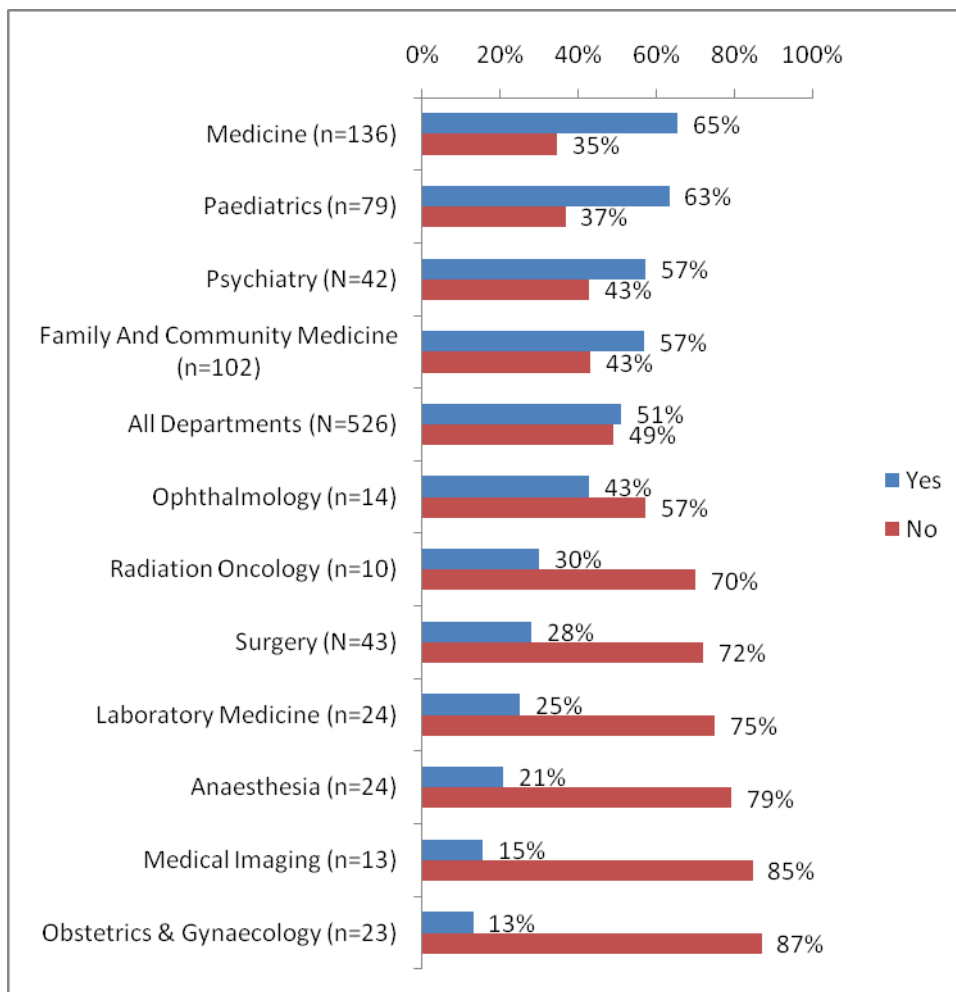
#### Background

In late 2008 through early 2009, the POWER Steering Committee released a survey to POWER users including trainees, teachers, and administrators. The survey asked users about several areas of POWER focusing on functions relating to evaluation. The following analyses present the results from teachers on the teacher evaluation function in POWER.

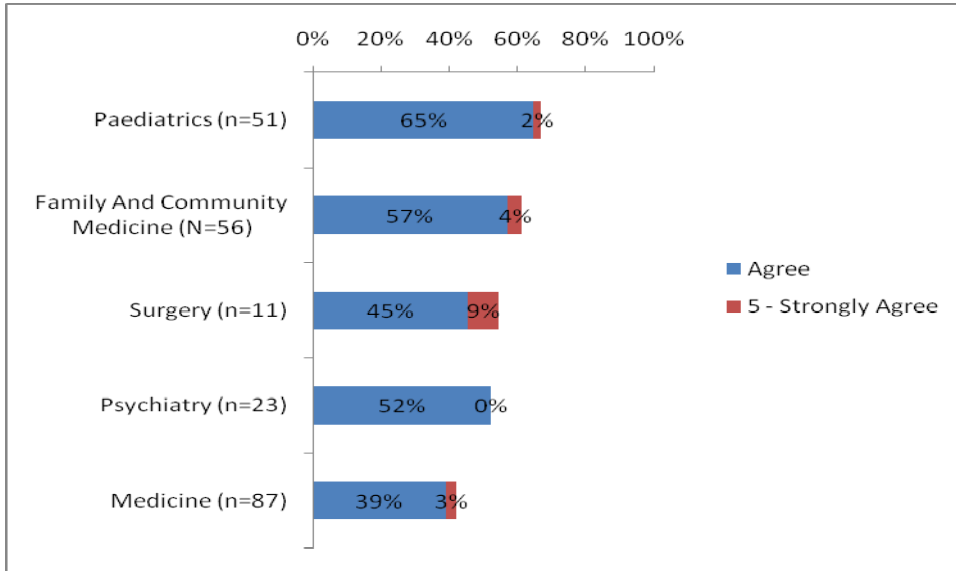
#### TES Question Responses by Department

Teachers were asked several specific questions regarding their experience with teacher evaluations in the POWER system. Survey results are presented below disaggregated by departmental affiliation of teachers.

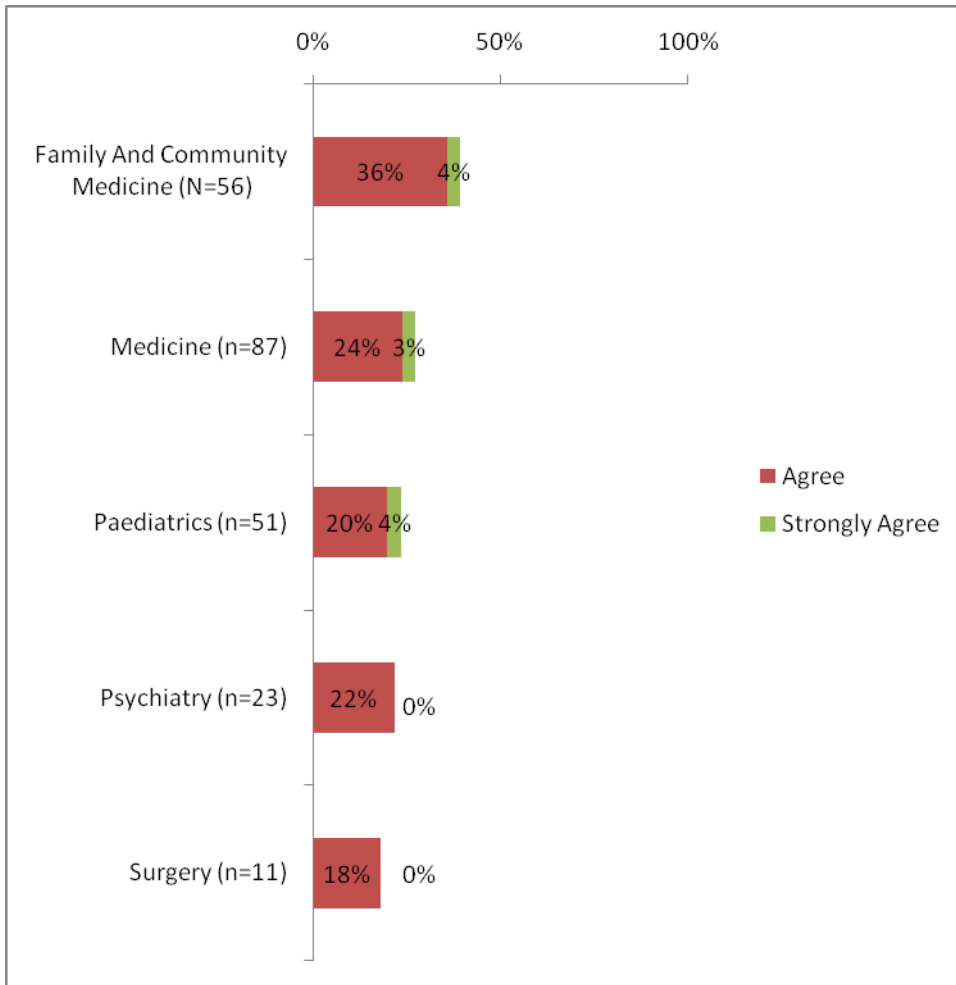
*I use POWER to review my TES online*



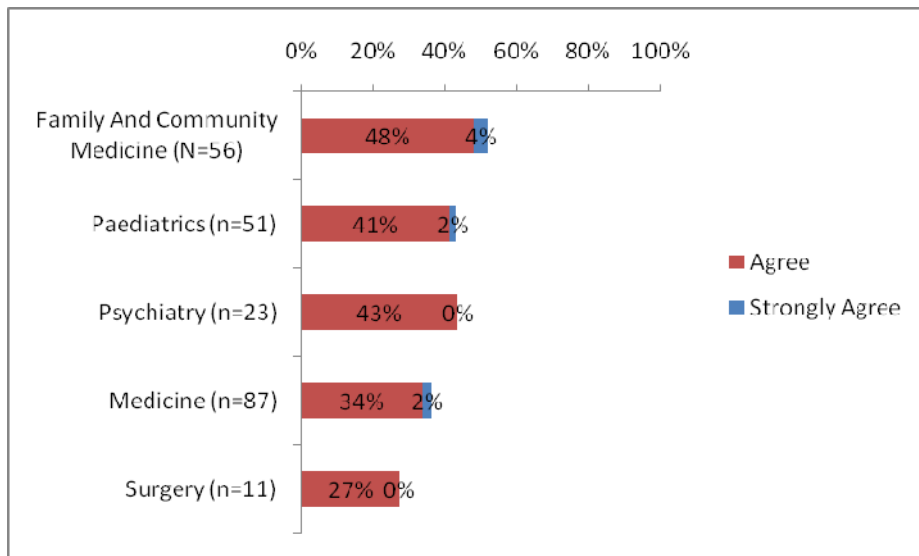
*TES is an appropriate tool to measure my performance and effectiveness as a teacher*



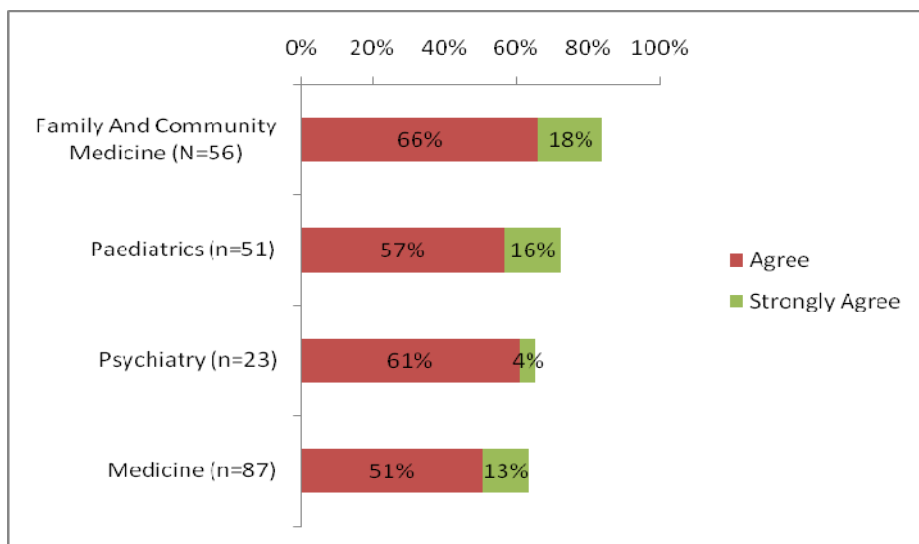
*Enough learners complete TES forms to give me an accurate and effective assessment of my performance*



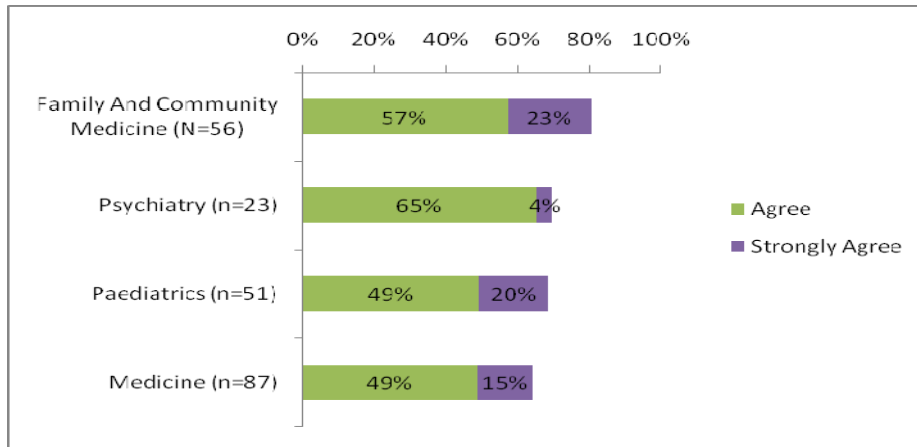
*The TES forms are current and relevant for my rotations and my learners*



*I find the qualitative comments useful and appropriate*



*I find the qualitative comments as useful as the quantitative ranking*



## Teacher Comments Relating to Teacher Evaluations

In analyzing the comments made by teachers in the 2009 POWER User Survey, a number of common themes emerged in comments regarding teaching evaluation.

### Qualitative Comments

- Enable comments for all questions/and or sections
- Require comments for very high or very low scores
- Comments are as useful or more useful than ratings
- Give residents training in constructive feedback via comments

### Evaluation Form Design

- Shorten the forms, and condense the domains evaluated
- Only have qualitative evaluation, no quantitative
- Use behavioural questions that do not ascribe motivation to actions
  - i.e. “Is on time for rounds” rather than “respect for trainees time”

### Confidentiality vs. Timeliness

- Holding forms back to protect trainees creates long feedback loop for teachers, too long to be useful
- Teachers would like to see their TES more frequently (every 3 or 6 months) regardless of how many are filled out
- Waiting for a minimum number of evaluations (usually 3) before teachers can view their scores leaves many without scores for > 1 year
- Trainees and teachers should be “blinded” from the results of each other’s evaluation until both are filled out

### Number of Evaluations Filled by Teachers

- Trainees should be required to fill out teacher evaluations
- Trainees should not get to see their ITERS until they have filled out a TES (to encourage more trainees to fill out evaluations)

### Appropriate Domains for Trainees to Assess

- Trainees should not assess the clinical skill, judgment, or management skills of a teacher
- Trainees should assess teaching ability only
- Early stage trainees (PGY1s) can be asked to assess the clinical skill of an attending, which they have no basis to accurately gauge.

### **Unfair/Inappropriate Evaluations**

- Trainees do not put enough thought into evaluations
- Trainees use teacher evaluations as a tool to retaliate for bad ITERS
- Anonymity encourages “glib” comments and evaluations
- A mechanism to view and address “outlier” TES scores is needed



## APPENDIX SIX: Relationships between TES and ITERs

### Correlations between ITER Scores and Teacher Evaluation Scores: *An Analysis prepared for the Best Practices in Teacher Assessment* *Working Group*

This report analyses the relationship between the scores given to residents on In Training Evaluation Reports (ITERs), and the scores this cohort of trainees gave their teachers on Teacher Evaluations. The data used are from an export of evaluation data from the Postgraduate Web Evaluation and Registration (POWER) system for the 2008-09 academic sessions. The report seeks to address 3 questions:

- 1) Do Trainees who receive a low ITER score from a teacher give those teachers lower teacher evaluations?
- 2) If trainees are required to fill out a teacher evaluation in order to see their ITER, does it affect the scores they give on teacher evaluations?
- 3) Do students who receive detailed feedback about their rotation performance give teachers higher or lower scores on their TE?

### **Methodology**

The data from POWER contain results of ITERs and teacher evaluations for each rotation. Only data from internal rotations were used, trainees on off-service rotations were excluded from the analysis. To analyse the effect of ITER scores on teacher evaluations, data was sorted by rotation so that the results of each ITER were paired with the teacher evaluations from the same rotation. The overall question from each form was used, and only ITERs and Teacher Evaluations using a 5 point overall question were included in the analysis.

ITER scores were sorted by overall score given (a discrete value from 1 to 5) and the average TES score was taken for each number on the ITER scale (e.g. the average TES score for students scoring a 1 on their ITER). In this way, a relationship between ITER scores, and TES scores was determined. A Wilcoxon Signed Rank Test was conducted to determine the significance (if any) of the results. The detailed findings are presented at the end of the report in Appendix A.

Other rotation characteristics such as the requirement for RES and TES forms to be completed before revealing an ITER to the trainee, and whether trainees received detailed feedback on their performance in a rotation were also used in the correlation analysis. When ITER results were separated based on these criteria, a Mann Whitney test was used to determine if the results of the two groups were significantly different. The detailed findings are presented at the end of the report in Appendix C.

To follow is an analysis of each of the questions asked in this report.

### **Do Trainees who receive lower ITER scores give lower teacher evaluations?**

ITER and TES results from all training programs conforming to the methodology described above were combined to determine if there was a relationship between the scores given to a trainee on an ITER, and the score that trainee gave to their teachers on that rotation.

According to the analysis, for each point an ITER drops across all residency programs, TES scores go down by 0.16 points.

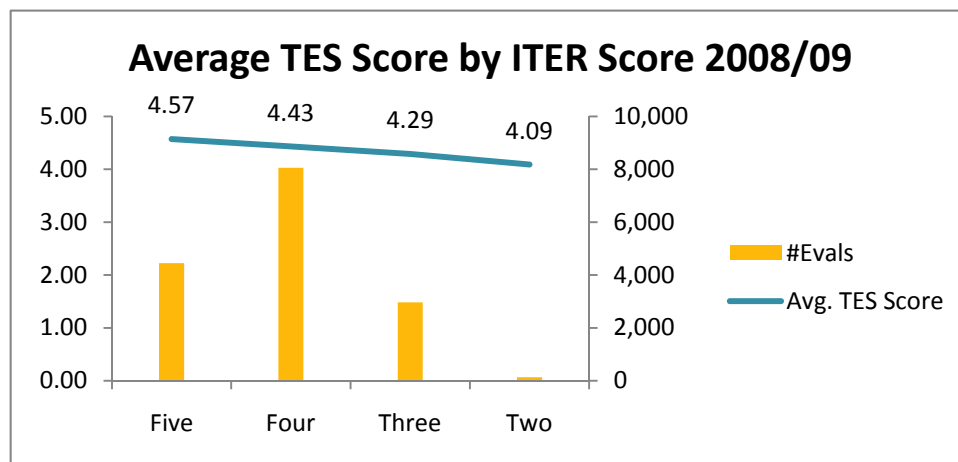
**Significance of Results**

A Wilcoxon Signed Rank Test was performed to test the hypothesis that learners with lower ITER scores in turn assign lower TES scores when assessing their teachers. The test showed a significant relationship between ITER and TES scores at the 0.01 level (p=0.000). The table below shows the direction of the relationship between the variables. In the context of a heavily top-weighted TES score distribution in which 90% of scores are 4 or 5, ITER scores can be said to correlate with TES scores strongly ( $r(15,605)=0.136$ ). For further background on the Wilcoxon Signed Rank Test, as well as the detailed analysis, please see Appendix A.

Similar analysis by major residency programs shows that the trend holds across programs.

**All Programs (2008-09)**

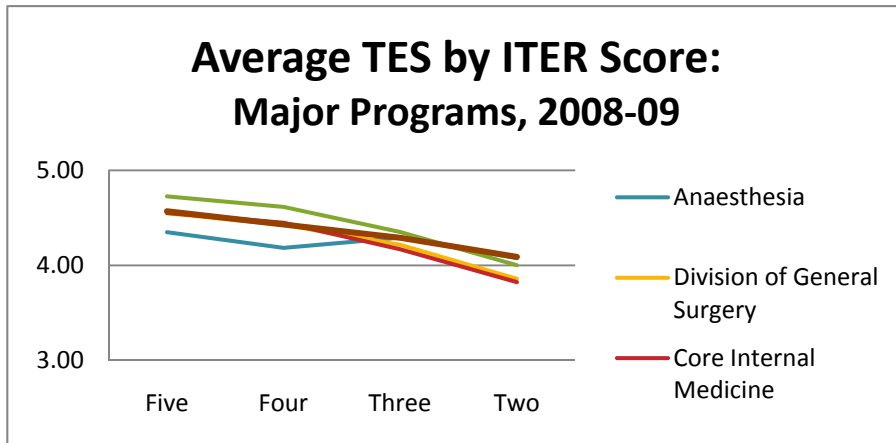
ITER Score	Avg. TES Score	#Evals
Five	4.57	4451
Four	4.43	8055
Three	4.29	2965
Two	4.09	134
<b>Total Mean</b>	<b>4.44</b>	<b>15605</b>



**Major Programs (2008-09)**

ITER	Anaesthesia TES	General Surgery TES	Internal Medicine TES	Paediatrics TES
------	-----------------	---------------------	-----------------------	-----------------

Five	4.35 (n= 996)	4.55 (n= 1319)	4.55 (n= 6771)	4.73 (n= 846)
Four	4.18 (n= 5432)	4.43 (n= 1908)	4.45 (n= 8166)	4.61 (n= 1172)
Three	4.29 (n= 2519)	4.22 (n= 430)	4.17 (n= 1113)	4.35 (n= 248)
Two	4.08 (n= 53)	3.86 (n= 27)	3.82 (n= 65)	



## Does the requirement to complete a teacher evaluation prior to receiving their ITER affect teacher evaluation scores?

All Programs in POWER have a setting available that allows for ITER results to be hidden from a trainee until at least one teacher evaluation and one rotation evaluation have been completed for the rotation. Programs were grouped by their use of this setting, and a full listing of which programs use this setting is in given in Appendix B.

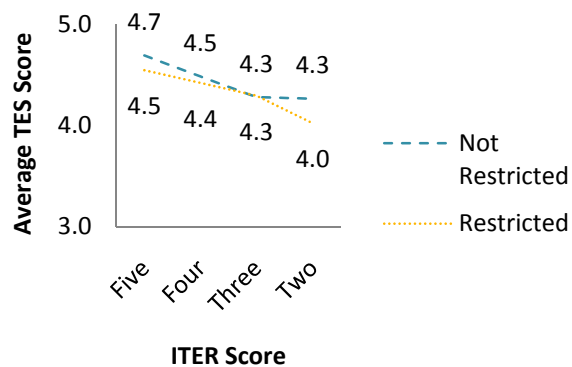
Teacher evaluation results from programs that use this setting and those that do not are presented below. The analysis shows that programs which do not use this setting see an average improvement of 0.11 pts on their TES scores.

### Significance of Results

A Mann-Whitney test was conducted to test the hypothesis that learners for whom ITERs were restricted based on having filled out rotation and teacher evaluations would assign different TES compared to those learners for whom ITERs were not restricted. The test showed a significant difference between the two groups ( $p=0.000$ ), with the group that was not required to fill out teacher evaluations before seeing their ITER tending to assign higher TES. For more background on the Mann-Whitney test, and detailed analysis, please see Appendix C.

**Average TES Results by Programs Requiring TE to be filled before a trainee can see their ITER**

ITER Score	Restricted: Avg. TES	Not Restricted: Avg. TES
Five	4.7	4.5
Four	4.5	4.4
Three	4.3	4.3
Two	4.3	4.0



## Do residents who receive detailed feedback about their rotation performance give higher or lower scores on their TE?

When students sign in to POWER to review their ITER forms, they are asked to agree or disagree with the following statement: “I received detailed verbal feedback on my performance at or near the end of the rotation”. Trainees who received detailed feedback are those who discuss their performance with their teacher at the end of their training, often before they complete their teacher evaluations. Trainee responses from the 2008-09 data were sorted according to their answer on this question.

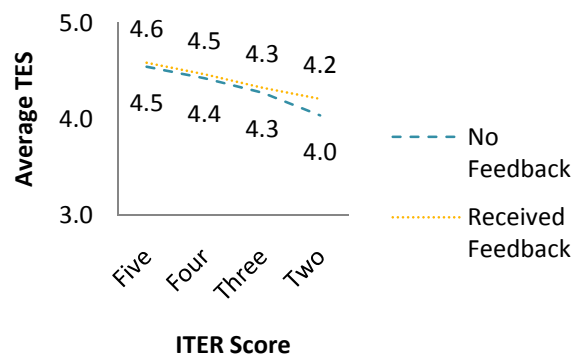
The analysis shows that trainees who responded yes to having detailed feedback rated their teachers higher than those who indicated that they did not receive feedback. Receiving feedback added an average of 0.07 points to a teacher evaluation, with the highest difference seen for low ITERs (score of 2) where feedback made a difference of 0.2 points.

### Significance of Results

To test the hypothesis that learners who received verbal reviews of their ITERs tended to provide higher TES, a Mann-Whitney test was conducted to compare the TES of learners who received verbal reviews to those that did not. The test showed a significant difference between the two groups ( $p=0.000$ ), with the group that received verbal reviews tending to assign higher TES. For more background on the Mann-Whitney test, and detailed analysis, please see Appendix C.

### Average TE Results by Trainees who did/did not receive feedback on their performance

ITER Score	No Feedback: Avg. TES	Received Feedback: Avg. TES
Five	4.5	4.6
Four	4.4	4.5
Three	4.3	4.3
Two	4.0	4.2



### Implications

The analysis shows that there is relativity between the ITER scores assigned by a teacher and the evaluation of that teacher by the learner. The mandatory completion of TES and/or RES in advance of receiving ITERs and the presence of verbal feedback also have an impact on the teacher evaluation scores.

The implications of these findings can be further discussed by the Best Practices in Teaching Assessment Working group.

## Appendix A – Statistical Validity of Link between ITERs and TES

### Background

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test for the case of two related samples or repeated measurements on a single sample. It can be used as an alternative to the paired Student's t-test when the population cannot be assumed to be normally distributed. The test is named for Frank Wilcoxon (1892–1965) who, in a single paper, proposed both it and the rank-sum test for two independent samples (Wilcoxon, 1945).

Like the paired or related sample t-test, the Wilcoxon test involves comparisons of differences between measurements, so it requires that the data are measured at an interval level of measurement. However it does not require assumptions about the form of the distribution of the measurements. It should therefore be used whenever the distributional assumptions that underlie the t-test cannot be satisfied. Findings

A Wilcoxon Signed Rank Test was performed to test the hypothesis that learners with lower ITER scores in turn assign lower TES scores when assessing their teachers. The test showed a significant relationship between ITER and TES scores at the 0.01 level ( $p=0.000$ ).

The table below shows the direction of the relationship between the variables. In the context of a heavily top-weighted TES score distribution in which 90% of scores are 4 or 5, ITER scores can be said to correlate with TES scores strongly ( $r(15,605)=0.136$ ).

<b>ITER Score</b>	<b>Avg. Score</b>	<b>TES</b>
Five	4.57	
Four	4.43	
Three	4.29	
Two	4.09	

<b>Score</b>	<b>ITER</b>	<b>TES</b>
1	0	68
2	134	227
3	2,964	1,327
4	8,055	5,120
5	4,451	8,862
Total	15,604	15,604
Mean	4.08	4.44
Std Dev	0.71	0.75

## Appendix B – Programs Restricting ITER Viewing until a Rotation and Teaching Evaluation are Filled

Program	ITERS after TES/RES	# Rotations
Anaesthesia	Y	224
Cardiology (Int Med)	Y	223
Core Internal Medicine	Y	2283
Critical Care Medicine	Y	77
Dermatology (Int Med)	Y	97
Division of Cardiac Surgery	Y	22
Division of Colorectal Surgery	Y	1
Division of General Surgery	Y	234
Division of Neurosurgery	Y	61
Division of Orthopaedic Surgery	Y	202
Division of Plastic Surgery	Y	46
Division of Thoracic Surgery	Y	13
Division of Urology	Y	71
Division of Vascular Surgery	Y	4
Emergency Medicine (Int Med)	Y	134
Endocrinology & Metabolism (Int Med)	Y	90
Family and Community Medicine	Y	3429
Family Medicine - Emergency Medicine	Y	25
Gastroenterology (Int Med)	Y	67
Geriatric Medicine (Int Med)	Y	44
Haematology (Int Med)	Y	65
Immunology & Allergy (Int Med)	Y	17
Infectious Disease (Int Med)	Y	18
Medical Oncology (Int Med)	Y	43
Neurology (Int Med)	Y	77
ObGyn - Maternal-Fetal Medicine Fellowship	Y	37
Obstetrics & Gynaecology	Y	261
Occupational Medicine (Int Med)	Y	6
Otolaryngology - Head and Neck Surgery	Y	90
Paediatric Emergency Medicine	Y	19
Paediatric Haematology/Oncology	Y	112
Paediatric Infectious Disease	Y	38
Paediatric Neurology	Y	59



Program	ITERS after TES/RES	# Rotations
Paediatric Respiratory Medicine	Y	51
Paediatric Rheumatology	Y	16
Palliative Medicine	Y	23
Physiatry (Int Med)	Y	83
Radiation Oncology	Y	202
Radiology - Diagnostic	Y	771
Radiology - Neuroradiology	Y	4
Respirology (Int Med)	Y	96
Rheumatology (Int Med)	Y	34
Community Medicine	N	7
General Internal Medicine (Int Med)	N	23
Laboratory Medicine Anatomical Pathology	N	279
Laboratory Medicine General Pathology	N	11
Laboratory Medicine Hematological Pathology	N	13
Laboratory Medicine Medical Microbiology	N	6
Medical Genetics	N	78
Nephrology (Int Med)	N	53
Ophthalmology	N	60
Paediatric Cardiology	N	27
Paediatric Clinical Immunology & Allergy	N	48
Paediatric Clinical Pharmacology	N	4
Paediatrics	N	627
Paediatrics - Developmental Paediatrics	N	42
Psychiatry	N	636

## Appendix C – Statistical Significance of Difference between ITER Results

### Background

In statistics, the Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) is a non-parametric test for assessing whether two independent samples of observations come from the same distribution. It is one of the best-known non-parametric significance tests. It was proposed initially by Frank Wilcoxon in 1945, for equal sample sizes, and extended to arbitrary sample sizes and in other ways by H. B. Mann and Whitney (1947). MWW is virtually identical to performing an ordinary parametric two-sample t test on the data after ranking over the combined samples.

### Restricted ITERs - Findings

A Mann-Whitney test was conducted to test the hypothesis that learners for whom ITERs were restricted based on having filled out rotation and teacher evaluations would assign different TES compared to those learners for whom ITERs were not restricted. The test showed a significant difference between the two groups ( $p=0.000$ ), with the group that was not required to fill out teacher evaluations before seeing their ITER tending to assign higher TES.

Rank Comparison		
Restricted ITERs?	N	Mean Rank
No	2,615	8,132.06
Yes	12,989	7,736.15
Total	15,604	

Test Statistics	
Mann-Whitney U	1.61E+07
Wilcoxin W	1.01E+08
Z	-4.64
Significance (2-tailed)	0.000

### Verbal Reviews of ITER Scores - Findings

To test the hypothesis that learners who received verbal reviews of their ITERS tended to provide higher TES, a Mann-Whitney test was conducted to compare the TES of learners who received verbal reviews to those that did not. The test showed a significant difference between the two groups ( $p=0.000$ ), with the group that received verbal reviews tending to assign higher TES.

Rank Comparison		
Verbal Review?	N	Mean Rank
No	3,808	6,351.76
Yes	9,325	6,654.9
Total	13,133	

Test Statistics	
Mann-Whitney U	1.69E+07
Wilcoxin W	2.42E+07
Z	-4.736
Significance (2-tailed)	0.000

## APPENDIX SEVEN: References

- Afonso, N. M., Cardozo, L. J., Mascarenhas, O. A., Aranha, A. N., & Shah, C. (2005). Are anonymous evaluations a better assessment of faculty teaching performance? A comparative analysis of open and anonymous evaluation processes. *Fam Med*, 37(1), 43-47.
- Al-Jarallah, K. F., Moussa, M. A., Shehab, D., & Abdella, N. (2005). Use of interaction cards to evaluate clinical performance. *Med Teach*, 27(4), 369-374.
- Ambrozy, D. M., Irby, D. M., Bowen, J. L., Burack, J. H., Carline, J. D., & Stritter, F. T. (1997). Role models' perceptions of themselves and their influence on students' specialty choices. *Acad Med*, 72(12), 1119-1121.
- Atzema C, B. G., Schull M. (2005). *Annals of Emergency Medicine*, 45(3), in press.
- Atzema C, B. G., Schull MJ. (2005). Emergency department crowding: The effect on resident education. *Annals of Emergency Medicine*, 45(3), in press.
- Austin, S. M., Balas, E. A., Mitchell, J. A., & Ewigman, B. G. (1994). Effect of physician reminders on preventive care: meta-analysis of randomized clinical trials. *Proc Annu Symp Comput Appl Med Care*, 121-124.
- Baker, M., & Sprackling, P. D. (1994). The educational component of senior house officer posts: differences in the perceptions of consultants and junior doctors. *Postgrad Med J*, 70(821), 198-202.
- Balestrieri, P. J. (2009). Validity and Reliability of Faculty Evaluations. *Anesthesia & Analgesia*, 108(6), 1991.
- Bandiera, G., & Regehr, G. (2003). Evaluation of a structured application assessment instrument for assessing applications to Canadian postgraduate training programs in emergency medicine. *Acad Emerg Med*, 10(6), 594-598.
- Bandiera, G., & Regehr, G. (2004). Reliability of a structured interview scoring instrument for a Canadian postgraduate emergency medicine training program. *Acad Emerg Med*, 11(1), 27-32.
- Bandiera G, L. S., Foote J. (2004). Effectiveness of a directed Faculty Development workshop for ED teaching. *8th annual University of Toronto Annual Faculty Research Day*.
- Bandiera G, L. S., Tiberius R. (2005). Creating effective learning in the emergency department: How expert teachers get it done. *Acad Emerg Med*, 45(3), in press.
- Bandiera G, L. S., Foote J. (2005). A faculty development workshop on emergency medicine teaching: timely and effective. *CJEM*, 7(2), xx-yy.
- Bandiera G, L. S., Foote J, Tiberius R. (2005). Perceptions and impact of a faculty development program for emergency department teaching. *Canadian Journal of Emergency Medicine*, Submitted.
- Bandiera G, T. L., Lee S, Tiberius R. (2004). What Emergency Medicine learners wish their teachers knew. *Clinical and Investigative medicine*, 27(4), 201.

- Bandiera, G. W., Morrison, L. J., & Regehr, G. (2002). Predictive validity of the Global Assessment Form used in a final-year undergraduate rotation in emergency medicine. *Acad Emerg Med*, 9(9), 889-895.
- Barnard, K., Elnicki, D. M., Lescisin, D. A., Tulsy, A., & Armistead, N. (2001). Students' perceptions of the effectiveness of interns' teaching during the internal medicine clerkship. *Acad Med*, 76(10 Suppl), S8-10.
- Barratt, M. S., & Moyer, V. A. (2004). Effect of a teaching skills program on faculty skills and confidence. *Ambul Pediatr*, 4(1 Suppl), 117-120.
- Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal of general internal medicine*, 20(12), 1159-1164.
- Beckman, T. J., Ghosh, A. K., Cook, D. A., Erwin, P. J., & Mandrekar, J. N. (2004). How reliable are assessments of clinical teaching? *Journal of general internal medicine*, 19(9), 971-977.
- Binder, L. S., & Chapman, D. M. (1995). Qualitative research methodologies in emergency medicine. *Acad Emerg Med*, 2(12), 1098-1102.
- Bordage, G., Burack, J. H., Irby, D. M., & Stritter, F. T. (1998). Education in ambulatory settings: developing valid measures of educational outcomes, and other research priorities. *Acad Med*, 73(7), 743-750.
- Borman, W. (1975). Effects of instruments to avoid halo error on reliability and validity of Performance evaluation ratings. *Journal of Applied Psychology*, 60(5), 556-560.
- Brennan, B. G., & Norman, G. R. (1997). Use of encounter cards for evaluation of residents in obstetrics. *Acad Med*, 72(10 Suppl 1), S43-44.
- Calkins, E. V., & Wakeford, R. (1983). Perceptions of instructors and students of instructors' roles. *J Med Educ*, 58(12), 967-969.
- Carter AJ, M. W. (2003). Off-service residents in the emergency department: the need for learner-centeredness. *Can Jnl Emerg Med*, 5(6), 400-405.
- Castle, N., Garton, H., & Kenward, G. (2007). Confidence vs competence: basic life support skills of health professionals. *Br J Nurs*, 16(11), 664-666.
- Chisholm, C. D., Collison, E. K., Nelson, D. R., & Cordell, W. H. (2000). Emergency department workplace interruptions: are emergency physicians "interrupt-driven" and "multitasking"? *Acad Emerg Med*, 7(11), 1239-1243.
- Chisholm, C. D., Dornfeld, A. M., Nelson, D. R., & Cordell, W. H. (2001). Work interrupted: a comparison of workplace interruptions in emergency departments and primary care offices. *Ann Emerg Med*, 38(2), 146-151.
- Coates WC , C. H., A Birnbaum , et al. (2003). Faculty development: academic opportunities for emergency medicine faculty on education career tracks. *Acad Emerg Med*, 10(10), 1113-1117.

- Coderre, S. P., Harasym, P., Mandin, H., & Fick, G. (2004). The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Med Educ*, 4, 23.
- Cohen, R., Rothman, A. I., Poldre, P., & Ross, J. (1991). Validity and generalizability of global ratings in an objective structured clinical examination. *Acad Med*, 66(9), 545-548.
- Cowan, J. A., Heckerling, P. S., & Parker, J. B. (1992). Effect of a fact sheet reminder on performance of the periodic health examination: a randomized controlled trial. *Am J Prev Med*, 8(2), 104-109.
- Crabtree, F. a. M., W.L. (1999). *Doing Qualitative Research* (2 ed.). Thousand Oaks, California: Sage Publications.
- D Swanson, G. W., J Norcini. (1990). Precision of patient ratings of residents' humanistic qualities: How many items and patients are enough? In D. Swanson (Ed.), *Teaching and assessing clinical competence* (pp. 424-431): Groningen.
- Davis, D. A., & Taylor-Vaisey, A. (1997). Translating guidelines into practice. A systematic review of theoretic concepts, practical experience and research evidence in the adoption of clinical practice guidelines. *Cmaj*, 157(4), 408-416.
- Davis, J. K., & Inamdar, S. (1990). The reliability of performance assessment during residency. *Acad Med*, 65(11), 716.
- Davis, J. K., Inamdar, S., & Stone, R. K. (1986). Interrater agreement and predictive validity of faculty ratings of pediatric residents. *J Med Educ*, 61(11), 901-905.
- Dawson-Saunders, B., & Paiva, R. E. (1986). The validity of clerkship performance evaluations. *Med Educ*, 20(3), 240-245.
- Devera-Sales, A., Paden, C., & Vinson, D. C. (1999). What do family medicine patients think about medical students' participation in their health care? *Acad Med*, 74(5), 550-552.
- Dillman, D. (1978). *Mail and Telephone Surveys: The complete method*. Troy, NY, USA: Wiley.
- Drescher, U., Warren, F., & Norton, K. (2004). Towards evidence-based practice in medical training: making evaluations more meaningful. *Med Educ*, 38(12), 1288-1294.
- Dudek, N. L., Marks, M. B., & Regehr, G. (2005). Failure to fail: the perspectives of clinical supervisors. *Acad Med*, 80(10 Suppl), S84-87.
- Durand, R. P., Levine, J. H., Lichtenstein, L. S., Fleming, G. A., & Ross, G. R. (1988). Teachers' perceptions concerning the relative values of personal and clinical characteristics and their influence on the assignment of students' clinical grades. *Med Educ*, 22(4), 335-341.
- Ende, J. (1983). Feedback in clinical medical education. *Jama*, 250(6), 777-781.
- Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: a reformulation and research agenda. *Acad Med*, 80(10 Suppl), S46-54.
- Eva, K. W., & Regehr, G. (2007). Knowing when to look it up: a new conception of self-assessment ability. *Acad Med*, 82(10 Suppl), S81-84.

- Farrell, S. E. (2005). Evaluation of student performance: clinical and professional performance. *Acad Emerg Med*, 12(4), 302e306-310.
- Ferenchick, G., Simpson, D., Blackman, J., DaRosa, D., & Dunnington, G. (1997). Strategies for efficient and effective teaching in the ambulatory care setting. *Acad Med*, 72(4), 277-280.
- Frank, J. R., & Langer, B. (2003). Collaboration, communication, management, and advocacy: teaching surgeons new skills through the CanMEDS Project. *World J Surg*, 27(8), 972-978; discussion 978.
- Furney, S. L., Orsini, A. N., Orsetti, K. E., Stern, D. T., Gruppen, L. D., & Irby, D. M. (2001). Teaching the one-minute preceptor. A randomized controlled trial. *J Gen Intern Med*, 16(9), 620-624.
- Gallagher EJ, M. M. (2000). *FACULTY DEVELOPMENT HANDBOOK*. Lansing, MI: Society for Academic Emergency Medicine.
- Gjerde, C. L., & Coble, R. J. (1982). Resident and faculty perceptions of effective clinical teaching in family practice. *J Fam Pract*, 14(2), 323-327.
- Gray, J. D. (1996). Global rating scales in residency education. *Acad Med*, 71(1 Suppl), S55-63.
- Griffin, A. (1989). Nurses' perceptions of their work environment. Organizational change on an adolescent unit. *Can J Nurs Adm*, 2(2), 19-23.
- Haber, R. J., & Avins, A. L. (1994). Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence? *J Gen Intern Med*, 9(3), 140-145.
- Hamilton, G. C., & Brown, J. E. (2003). Faculty development: what is faculty development? *Acad Emerg Med*, 10(12), 1334-1336.
- Heidenreich, C., Lye, P., Simpson, D., & Lourich, M. (2000). The search for effective and efficient ambulatory teaching methods through the literature. *Pediatrics*, 105(1 Pt 3), 231-237.
- Hodges, B. D. (2006). The Objective Structured Clinical Examination: three decades of development. *J Vet Med Educ*, 33(4), 571-577.
- Irby, D. M. (1978). Clinical teacher effectiveness in medicine. *J Med Educ*, 53(10), 808-815.
- Irby, D. M. (1994). What clinical teachers in medicine need to know. *Acad Med*, 69(5), 333-342.
- Irby, D. M. (1995). Teaching and learning in ambulatory care settings: a thematic review of the literature. *Acad Med*, 70(10), 898-931.
- Irby, D. M., Aagaard, E., & Teherani, A. (2004). Teaching points identified by preceptors observing one-minute preceptor and traditional preceptor encounters. *Acad Med*, 79(1), 50-55.
- Irby, D. M., Ramsey, P. G., Gillmore, G. M., & Schaad, D. (1991). Characteristics of effective clinical teachers of ambulatory care medicine. *Acad Med*, 66(1), 54-55.



- JCadwell. (1986). Teachers' judgements about their students: The effects of cognitive simplification strategies on the rating process. *American Education research Journal*, 23(3), 460-475.
- Jouriles, N. J., Emerman, C. L., & Cydulka, R. K. (2002). Direct observation for assessing emergency medicine core competencies: interpersonal skills. *Acad Emerg Med*, 9(11), 1338-1341.
- Kendrick, S. B., Simmons, J. M., Richards, B. F., & Roberge, L. P. (1993). Residents' perceptions of their teachers: facilitative behaviour and the learning value of rotations. *Med Educ*, 27(1), 55-61.
- Kernan, W. N., Lee, M. Y., Stone, S. L., Freudigman, K. A., & O'Connor, P. G. (2000). Effective teaching for preceptors of ambulatory care: a survey of medical students. *Am J Med*, 108(6), 499-502.
- Kim, S., Kogan, J. R., Bellini, L. M., & Shea, J. A. (2005). A randomized-controlled study of encounter cards to improve oral case presentation skills of medical students. *J Gen Intern Med*, 20(8), 743-747.
- Kotzabassaki, S., Panou, M., Dimou, F., Karabagli, A., Koutsopoulou, B., & Ikononou, U. (1997). Nursing students' and faculty's perceptions of the characteristics of 'best' and 'worst' clinical teachers: a replication study. *J Adv Nurs*, 26(4), 817-824.
- Laidley, T. L., & Braddock, I. C. (2000). Role of Adult Learning Theory in Evaluating and Designing Strategies for Teaching Residents in Ambulatory Settings. *Adv Health Sci Educ Theory Pract*, 5(1), 43-54.
- Lee, W. S., Cholowski, K., & Williams, A. K. (2002). Nursing students' and clinical educators' perceptions of characteristics of effective clinical educators in an Australian university school of nursing. *J Adv Nurs*, 39(5), 412-420.
- Litzelman, D. K., Stratos, G. A., & Skeff, K. M. (1994). The effect of a clinical teaching retreat on residents' teaching skills. *Acad Med*, 69(5), 433-434.
- Loomis, J. (1985). Evaluating clinical competence of physical therapy students. Part 2: assessing the reliability, validity and usability of a new instrument. *Physiother Can*, 37(2), 91-98.
- Luker, K., Beaver, K., Austin, L., & Leinster, S. J. (2000). An evaluation of information cards as a means of improving communication between hospital and primary care for women with breast cancer. *J Adv Nurs*, 31(5), 1174-1182.
- Magill, M. K., McClure, C., & Commerford, K. (1986). A system for evaluating teaching in the ambulatory setting. *Fam Med*, 18(3), 173-174.
- Mahan, J. M., & Shellenberger, S. (1983). Changes in students' perceptions of clinical teaching as a result of general-practice clerkships. *Med Educ*, 17(3), 155-158.
- Maloney, J. P., Bartz, C., & Allanach, B. C. (1991). Staff perceptions of their work environment before and six months after an organizational change. *Mil Med*, 156(2), 86-92.
- McGee, S. R., & Irby, D. M. (1997). Teaching in the outpatient clinic. Practical tips. *J Gen Intern Med*, 12 Suppl 2, S34-40.

- McLeod, P. J., James, C. A., & Abrahamowicz, M. (1993). Clinical tutor evaluation: a 5-year study by students on an in-patient service and residents in an ambulatory care clinic. *Med Educ*, 27(1), 48-54.
- Miles MB, H. A. (1994). *Qualitative Data Analysis*. Thousand Oaks, CA: Sage.
- Morgan, J., & Knox, J. E. (1987). Characteristics of 'best' and 'worst' clinical teachers as perceived by university nursing faculty and students. *J Adv Nurs*, 12(3), 331-337.
- Neher, J. O., Gordon, K. C., Meyer, B., & Stevens, N. (1992). A five-step "microskills" model of clinical teaching. *J Am Board Fam Pract*, 5(4), 419-424.
- Noel, G. L., Herbers, J. E., Jr., Caplow, M. P., Cooper, G. S., Pangaro, L. N., & Harvey, J. (1992). How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med*, 117(9), 757-765.
- Norman, G. (1985). Defining Competence. In Neufeld (Ed.), *Assessing Clinical Competence* (pp. 15-13). New York: Springer Verlag.
- Paukert, J. L., Richards, M. L., & Olney, C. (2002). An encounter card system for increasing feedback to students. *Am J Surg*, 183(3), 300-304.
- Penciner, R. (2002). Clinical teaching in a busy emergency department: strategies for success. *ACn. Jnl. Emerg. Med.*, 4(4), 286-288.
- Pinsky, L. E., & Irby, D. M. (1997). "If at first you don't succeed": using failure to improve teaching. *Acad Med*, 72(11), 973-976; discussion 972.
- Probst, J. C., Baxley, E. G., Schell, B. J., Cleghorn, G. D., & Bogdewic, S. P. (1998). Organizational environment and perceptions of teaching quality in seven South Carolina family medicine residency programs. *Acad Med*, 73(8), 887-893.
- Ramalanjaona, G. (2003). Faculty development: how to evaluate the effectiveness of a faculty development program in emergency medicine. *Acad Emerg Med*, 10(8), 891-892.
- Ramani, S. (2003). Twelve tips to improve bedside teaching. *Med Teach*, 25(2), 112-115.
- Ramani, S., Orlander, J. D., Strunin, L., & Barber, T. W. (2003). Whither bedside teaching? A focus-group study of clinical teachers. *Acad Med*, 78(4), 384-390.
- Ramsbottom-Lucier, M. T., Gillmore, G. M., Irby, D. M., & Ramsey, P. G. (1994). Evaluation of clinical teaching by general internal medicine faculty in outpatient and inpatient settings. *Acad Med*, 69(2), 152-154.
- Ramsey, P. G., Gillmore, G. M., & Irby, D. M. (1988). Evaluating clinical teaching in the medicine clerkship: relationship of instructor experience and training setting to ratings of teaching effectiveness. *J Gen Intern Med*, 3(4), 351-355.
- Regehr, G., & Eva, K. (2006). Self-assessment, self-direction, and the self-regulating professional. *Clin Orthop Relat Res*, 449, 34-38.
- Reznick, R. (1991). In-training evaluation - it's more than just a form. *Annals of the Royal College of Physicians and Surgeons of Canada*, 24(6), 415-420.

- Ringsted, C., Ostergaard, D., Ravn, L., Pedersen, J. A., Berlac, P. A., & van der Vleuten, C. P. (2003). A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. *Med Teach, 25*(6), 654-658.
- Ross L, N. R. (1991). *The person and the situation - perspectives of social psychology*. Philadelphia: Temple University Press.
- Rubin HJ, R. I. (1995). Topical interviewing. In *Qualitative Interviewing: the art of hearing data* (pp. 196-225.). Thousand Oaks: Sage.
- Rubin HJ, R. I. (1995). What did you hear? In *Qualitative Interviewing: The art of hearing data* (pp. 226-256). Thousand Oaks: Sage.
- Ryan, J. G., Mandel, F. S., Sama, A., & Ward, M. F. (1996). Reliability of faculty clinical evaluations of non-emergency medicine residents during emergency department rotations. *Acad Emerg Med, 3*(12), 1124-1130.
- Scheuneman, A. L., Carley, J. P., & Baker, W. H. (1994). Residency evaluations. Are they worth the effort? *Arch Surg, 129*(10), 1067-1073.
- Silverman, D. (2001). Analyzing Talk and Text. In L. Y. Denzin N (Ed.), *The handbook of qualitative research* (pp. 821-834). Thousand Oaks: Sage.
- Skeff, K. (1997). Clinical teaching improvement: past and future for faculty development. *Fam Med, 29*, 252-257.
- Skeff, K. M. (1988). Enhancing teaching effectiveness and vitality in the ambulatory setting. *J Gen Intern Med, 3*(2 Suppl), S26-33.
- Skeff, K. M. (1995). Irby DM Teaching and learning in ambulatory care settings: a thematic review of the literature. *Acad Med. 1995; 70*:898-931.
- Skeff, K. M., Bowen, J. L., & Irby, D. M. (1997). Protecting time for teaching in the ambulatory care setting. *Acad Med, 72*(8), 694-697; discussion 693.
- Skeff, K. M., & Mutha, S. (1998). Role models--guiding the future of medicine. *N Engl J Med, 339*(27), 2015-2017.
- Skeff, K. M., Stratos, G. A., Bergen, M. R., Sampson, K., & Deutsch, S. L. (1999). Regional teaching improvement programs for community-based teachers. *Am J Med, 106*(1), 76-80.
- Skeff, K. M., Stratos, G. A., Mygdal, W., DeWitt, T. A., Manfred, L., Quirk, M., et al. (1997). Faculty development. A resource for clinical teachers. *J Gen Intern Med, 12 Suppl 2*, S56-63.
- Skeff, K. M., Stratos, G. A., Mygdal, W. K., DeWitt, T. G., Manfred, L. M., Quirk, M. E., et al. (1997). Clinical teaching improvement: past and future for faculty development. *Fam Med, 29*(4), 252-257.
- Smith, C. A., Varkey, A. B., Evans, A. T., & Reilly, B. M. (2004). Evaluating the performance of inpatient attending physicians. *Journal of general internal medicine, 19*(7), 766-771.
- Smith, C. S., & Irby, D. M. (1997). The roles of experience and reflection in ambulatory care education. *Acad Med, 72*(1), 32-35.

- Snell, L., Tallett, S., Haist, S., Hays, R., Norcini, J., Prince, K., et al. (2001). A review of the evaluation of clinical teaching: new perspectives and challenges\*. *Medical education*, 34(10), 862-870.
- SNWG. (1996). Skills for the new millennium. *Annals RCPSC*, 29(4), 206-216.
- Steiner, I. P., Franc-Law, J., Kelly, K. D., & Rowe, B. H. (2000). Faculty evaluation by residents in an emergency medicine program: a new evaluation instrument. *Acad Emerg Med*, 7(9), 1015-1021.
- Steiner, I. P., Yoon, P. W., Kelly, K. D., Diner, B. M., Blitz, S., Donoff, M. G., et al. (2005). The influence of residents training level on their evaluation of clinical teaching faculty. *Teach Learn Med*, 17(1), 42-48.
- Steiner, I. P., Yoon, P. W., Kelly, K. D., Diner, B. M., Donoff, M. G., Mackey, D. S., et al. (2003). Resident evaluation of clinical teachers based on teachers' certification. *Acad Emerg Med*, 10(7), 731-737.
- Steiner, J. F., Lanphear, B. P., Curtis, P., & Vu, K. O. (2002). The training and career paths of fellows in the National Research Service Award (NRSA) Program for Research in Primary Medical Care. *Acad Med*, 77(7), 712-718.
- Steinert Y, N. L., DAigle N. (2000). A Faculty Development Workshop on "Developing Successful Workshops". *Acad Med*, 75, 554-555.
- Streiner. (1985). Global Rating Scales. In N. V (Ed.), *Assessing Clinical Competence* (pp. 119-141). New York: Springer Verlag.
- Stritter, F. T., & Baker, R. M. (1982). Resident preferences for the clinical teaching of ambulatory care. *J Med Educ*, 57(1), 33-41.
- Taylor, A. S. (1994). Emergency medicine educational objectives for the undifferentiated physician. *J Emerg Med*, 12(2), 255-262.
- Ten Eyck, R. P., & Maclean, T. A. (1994). Improving the quality of emergency medicine rotation/clerkship evaluations. *Am J Emerg Med*, 12(1), 113-117.
- Tomic, V., Sporis, G., Nizic, D., & Galinovic, I. (2007). Self-reported confidence, attitudes and skills in practical procedures among medical students: questionnaire study. *Coll Antropol*, 31(3), 683-688.
- Ullian, J. A., Bland, C. J., & Simpson, D. E. (1994). An alternative approach to defining the role of the clinical teacher. *Acad Med*, 69(10), 832-838.
- Usatine, R. P., Nguyen, K., Randall, J., & Irby, D. M. (1997). Four exemplary preceptors' strategies for efficient teaching in managed care settings. *Acad Med*, 72(9), 766-769.
- Verma, S., Flynn, L., & Seguin, R. (2005). Faculty's and residents' perceptions of teaching and evaluating the role of health advocate: a study at one Canadian university. *Acad Med*, 80(1), 103-108.
- Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: current state of the art. *Adv Health Sci Educ Theory Pract*, 7(1), 63-80.
- Ward, M., MacRae, H., Schlachta, C., Mamazza, J., Poulin, E., Reznick, R., et al. (2003). Resident self-assessment of operative performance. *Am J Surg*, 185(6), 521-524.

- Weaver, C. S., Humbert, A. J., Besinger, B. R., Graber, J. A., & Brizendine, E. J. (2007). A more explicit grading scale decreases grade inflation in a clinical clerkship. *Acad Emerg Med*, 14(3), 283-286.
- White, C. B., Bassali, R. W., & Heery, L. B. (1997). Teaching residents to teach. An instructional program for training pediatric residents to precept third-year medical students in the ambulatory clinic. *Arch Pediatr Adolesc Med*, 151(7), 730-735.
- Wilkerson, L. (1998). Strategies for improving teaching practices: a comprehensive approach to faculty development. *Acad Med*, 73, 387-396.
- Wolverton, S. E., & Bosworth, M. F. (1985). A survey of resident perceptions of effective teaching behaviors. *Fam Med*, 17(3), 106-108.
- Woolliscroft, J. O., Howell, J. D., Patel, B. P., & Swanson, D. B. (1994). Resident-patient interactions: the humanistic qualities of internal medicine residents assessed by patients, attending physicians, program supervisors, and nurses. *Acad Med*, 69(3), 216-224.
- Woolliscroft, J. O., & Schwenk, T. L. (1989). Teaching and learning in the ambulatory setting. *Acad Med*, 64(11), 644-648.
- Xu, G., Wolfson, P., Robeson, M., Rodgers, J. F., Veloski, J. J., & Brigham, T. P. (1998). Students' satisfaction and perceptions of attending physicians' and residents' teaching role. *Am J Surg*, 176(1), 46-48.