SHE

PhD project proposal

Registration form

1a. Details of applicant

-Name, title(s): Shiphra Ginsburg MD, MEd
-Male/female: Female
-University, Department (or Institute): University of Toronto, Department of Medicine and Wilson Centre for Research in Education
-Address for correspondence: Mount Sinai Hospital; 433-600 University Ave, Toronto, ON M5G1X5
-Telephone: 416-586-8671
-E-mail: shiphra.ginsburg@utoronto.ca

1b. Title of research proposal

Towards a deeper understanding of evaluator subjectivity: An in-depth exploration of the language used in assessments

1c. Abstract

Obligatory!, max. 100 words

Subjectivity in assessment is gaining increasing respect in the medical education community. The overall goal of this proposed research program is to gain a deeper understanding of how clinical supervisors use their subjective impressions in the assessment of learners by analyzing the language they use in assessments. A secondary goal is to develop novel methods for using language analysis as a distinct part of assessment in a way that will be credible, valid and defensible.

1d. MSc (date and field) main applicant: Master of Education, Specialization in the Health Professions; Ontario Institute for Studies in Education, University of Toronto, June 1999

1e. Complete name dissertation supervisor(s)

If already known, please state the complete name of the dissertation supervisor(s) for the proposed research.

Prof. Cees van der Vleuten
Prof. Kevin Eva

Research proposal

3. Description of the proposed research

max. 4.000 words (excluding references, including footnotes) for 3a and 3b.
(use Word Count to specify number of words): 3976

3a. Research topic (theoretical framework, research questions, hypotheses)

Evaluating the clinical competence of residents remains a challenge for educators. Despite years of researchers and educators calling for instruments with ever higher degrees of objectivity and standardization, most residency programs still rely on in-training evaluation reports (ITERs) for a large portion of their assessments.[1] ITER-type forms are ubiquitous and familiar, yet have received much criticism over time for being "too subjective" and for suffering from poor reliability and validity. However, recent work has questioned these assumptions[2] and some have suggested that we are now entering an era of decline of the "dominance of psychometric discourse and a rise in discourses anchored in subjectivity"[3]

The case for subjectivity:
In an intriguing large-scale study Crossley et al showed that aligning response scales on evaluations with "the priorities of clinician assessors" significantly reduced assessor disagreement and improved reliability of 3 different assessment tools.[4] The authors further noted that each discipline likely has a different conception of what may be important in assessing its trainees so consequently each instrument should be unique. What is important is the "reality map", or "cognitive structure" of the assessors in each.[5] This suggests that, rather than attempting to make assessments "objective," instead we should strive to better understand influences on (and information derived from) subjectivity such that the added value subjectivity might provide can be effectively built into meaningful assessments. Hodges' work on OSCE exams suggested this many years ago by demonstrating that global ratings captured expertise far better than did a checklist approach.[6] He has recently argued that if we consider evaluation as being similar to clinical diagnosis, we can quickly see useful parallels – there is great value in the pattern recognition that experts use for diagnosis when compared to the checklist approaches more common to novices. [3] Indeed, the "entrustability" movement relies on supervisors to know when they can trust trainees to perform a given task or role independently.[7] These data together suggest that evaluators' gestalt or gut reactions can have great value, and that we need to tap into what supervisors actually perceive as important in terms of development of competence or expertise. One way to tap into how supervisors think is to explore the language they use in assessment.

The case for the utility of analyzing narrative comments in assessment:
In one early study, faculty who were blinded to medical students' numeric scores evaluated the narrative comments on their evaluation forms.[8] They identified significantly more students as being in difficulty when compared to students' tutors. It turned out that negative or problematic comments were actually present on those students' evaluations, but they were not reflected in the assigned scores. The authors concluded that evaluating the comments on evaluation forms and not just looking at the numbers would be beneficial to students' progress. In a similar study in the postgraduate setting, Durning et al reported an analysis of over 1500 Internal Medicine (IM) residents' evaluation forms collected over a 10 year period.[9] Although most comments were concordant with the numeric scores assigned, on a significant minority the comments were coded as more negative than the scores would suggest, particularly in the domains of attitude and clinical skills. Again, the authors concluded that the comments can reveal areas requiring attention and remediation that would otherwise have been missed if only focusing on the numeric scores.

In related work my research team conducted two studies in which faculty were asked to categorize and rank-order narrative profiles of IM residents. The first study used one-page profiles we created based on data from interviews with faculty in which they described outstanding and problematic residents they had worked with[10] and the second used comments compiled from current IM residents' ITERs.[2] Both studies found that faculty could easily and very reliably rank-order residents based on the comments alone (e.g., 3-rater reliability in study 2 was 0.83). In a third study we conducted a qualitative analysis of the narrative comments on 1770 IM residents' ITERs using constructivist grounded theory and found that assessors' language reflected many important "non-competency" factors including a resident's apparent "trajectory", which were otherwise not captured on the ITER form.[11] This suggested that evaluations should attempt to capture these integrated, holistic impressions, perhaps by analyzing the language more closely.

Thus, evidence suggests that the numeric ratings found on ITER forms may not be the most valuable aspect of these assessment tools, whereas conducting in-depth qualitative analyses of comments can provide very useful information that is potentially more educationally relevant to both trainees and faculty. However, such analyses are very labour-intensive. In response, some researchers have investigated the utility of "scoring" qualitative comments according to their positive/negative polarity (what some have called valence).

In one study of the evaluation of professionalism, for example, investigators found that the number of "positive" written comments on a student's evaluation form correlated with students' numerical professionalism score while the number of "negative" comments, although uncommon, had a negative correlation.[12] It is worth noting that the number of "equivocal" comments also had a significantly negative correlation; that is, they behaved exactly like negative comments. This suggests that it might be fruitful to look for these equivocal or lukewarm comments as they may indicate more significant issues for the student.

However, the actual process of scoring comments for polarity can be problematic. For example, in Frohna's study comments were coded as "equivocal" if they were neither wholly positive or wholly negative, and as a result these comments often contained both positive and negative phrases. The authors were clear that these comments were not, in fact, "neutral". To avoid the difficulties of scoring such complex comments, Canavan et al parsed multiple or complex statements into "feedback phrases" with a single point of focus, so that each could be scored as either positive or negative.[13] However this did not entirely solve the problem either as polarity could not be assessed for up to 20% of comments.

In none of the described studies was the *degree* of positivity or negativity of comments noted. That is, comments were either positive, negative, equivocal or neutral, but no effort was made to codify *how* positive or *how* negative. The most likely reason for this (personal communication with several authors) is that it can be difficult to agree on how positive or negative a comment is. It might be easy to agree that adjectives describing knowledge base such as excellent, outstanding or superb might reflect the same underlying meaning, but it may be more problematic to quantify the polarity of phrases like decent, solid, adequate, or appropriate. These are all "positive" but clearly less so and they can even be interpreted negatively depending on context (e.g., by providing lukewarm praise).[14] In fact, recent research has shown that such adjectives may have different connotations depending on which domains of performance on being assessed.[15] If these were scored differently it is unclear what effect that might have on the relationships noted thus far between comments and scores. These more nuanced interpretations may be obscured by simplified coding schema.

Related to this, there are limited data suggesting the importance of analyzing complete sentences rather than parsing into phrases. For example, in Frohna's study if those "equivocal" phrases were split into the wholly positive and wholly negative components the authors would have lost their ability to determine the overall effect of these complex phrases. In addition, Bogo et al found that when faculty social workers described students' problem areas they often used "'but' statements." For example, stating that a student was liked by the staff *but* was too casual and had poor boundaries with clients.[16] Similarly, in the interview study by my research team mentioned above, we found that experienced faculty attendings would often dismiss areas of strength in residents who were problematic, whereas they would discount or excuse deficiencies in residents who they thought were outstanding.[17] For the problematic residents, these areas of strength were often phrased as "but statements", as faculty often began their descriptions with a positive comment followed by a "but…". These phrases may be important indicators of faculty's opinions, but will not be captured well either by coding the statement as "equivocal" or by segmenting the sentence into two separate, equally weighted "feedback phrases".

In summary, then, evidence suggests that the narrative comments on assessments can provide important additional sources of assessment data, above and beyond numeric scores, perhaps even improving the ability of ITERs to identify residents in difficulty. Yet in practice comments are rarely used in a systematic way, as it is labour-intensive to analyze comments. In addition, it is not clear what the comments might add over and above the scores, and there is uncertainty about what to do if the scores and comments are discordant. Finally, the education community in general has been reluctant to embrace "subjective" means of assessment, greatly preferring quantitative and more apparently "objective" methods.

This raises two important research questions:

1. How can we use the language from assessments to deepen our understanding of how faculty conceptualize competence and performance?
2. Can we create methods for using language analysis as a distinct part of assessment in a way that will be credible, valid and defensible?

Analysis of language by software programs:
The analysis of language is complex and has a long history, a full exploration of which is beyond the scope of this proposal. The construct of "affective language" appears to be important, however, in representing the emotional "tone" of written or spoken language. Several computer programs have been developed to analyze affective language, by accepting a phrase, sentence or paragraph of text, and almost immediately producing a set of scores based on the words comprising the text. The most relevant for the purposes of this proposal appear to be the Dictionary of Affect in Language (DAL)[18] and the Linguistic Inquiry and Word Count (LIWC; www.liwc.net).

The DAL was developed in 1989 to quantify the "undertones (connotations, associations) of emotional words". [18] The corpus, or list of words it contains, has been extensively refined and validated over many years, and now contains over 8700 words, capturing over 90% of natural language. The current version analyzes text on three core dimensions: Pleasantness (on a 3-point scale of unpleasant – in between – pleasant), Activation (on a 3-point scale of passive – in between – active) and Imagery/concreteness (on a 3-point scale of difficult to envision – in between – easy to envision). Each word has been scored for each of the 3 dimensions, in a process involving more than 200,000 individual rating judgments. A word's score consists of the means of volunteers' ratings of that word on each dimension. Samples of language entered into the software are scored in a word-by-word matching procedure. When a match is found in the corpus the program imports the values for Pleasantness, Activation and Imagery for that word into a data file. The mean for all values characterizes the sample as a whole. For example, the sentence "Glad to see you" would be scored at 2.25(P), 1.89(A), 1.55(I), so can be characterized as showing high pleasantness, moderately high activation, and average imagery when compared with everyday language. The DAL has been shown capable of differentiating between the emotional tone and affect of "voices" in many different settings.[18] However, in interpreting the scores from the DAL, it is important to note Whissell's advice that it is most appropriate to utilize it "in conjunction with theories characterizing the domains in which it is applied."[18]

The LIWC was developed by Pennebaker and colleagues in the mid-1990's as a way to provide an efficient and effective method for studying the emotional, cognitive and structural components present in verbal and written speech. It analyzes text to determine the rate at which writers use positive or negative emotion words, as well as approximately 80 other dimensions of word use. Similar to the DAL, entered text is matched on a word-by-word basis to entries in the program. If a match is found the appropriate categories and scales for that word are output in a data file which includes some general descriptor categories such as total word count, and then looks at 22 standard linguistic dimensions (e.g., percentage of words that are pronouns, verbs, etc), 32 categories that tap into psychological constructs such as affect and cognition, and several other categorizations including punctuation, etc. Once found, each word is matched to applicable categories in a hierarchical way, so that if a word matches for negative emotion it also counts as overall affect. As an example, the word "cried" would fall under five categories: sadness, negative emotion, overall affect, verb, past tense verb. The LIWC has been shown to have good internal reliability and external validity in many studies involving written and spoken language samples.[19] In one typical study expert judges rated essays on various emotional and cognitive levels and their ratings correlated highly with similar dimensions on the LIWC.[20] The use of pronouns in written language has proved particularly fascinating.[21] Again, however, as with the DAL, this has not been used or tested in medical education assessment settings.

Other relevant work to date:
In addition to the studies already described, we have been exploring the potential utility of these two software programs. To start, we first attempted to assess the ability of both scores and comments from PGY1 and PGY2 ITERs to predict PGY3 performance in our IM residency program.[2] To analyze the comments we had 24 faculty, blinded as to resident or evaluator identity, rank and sort comments that were extracted from the residents' ITERs. Each resident (approx. 60/yr) had one document made up of PGY1 comments and one from PGY2 comments. Faculty participants were able to rank and sort the comments with a high degree of reliability (0.83 for 3 raters per resident), similar to our "crafted" profiles mentioned above.[10] However, the ITER scores were actually fairly reliable on their own (0.59), so

adding in the comment scores did not add to the overall predictive model. That said, the high reliability in ranking and sorting the comments suggests that faculty have a shared interpretation of the language used on the evaluations, reflecting perhaps a shared model of resident performance and competence.

Preliminary work using the LIWC and DAL on these data was promising. We first tested the 16 "crafted" resident profiles[17] with relevant measures from the LIWC and found good correlations with the rank-order of the profiles generated by faculty, e.g., r= 0.63 for "negative emotions" in the lower ranked profiles and r=0.89 for "positive emotions" in the higher ranked profiles. In further testing of the LIWC we selected 14 of the 80 available variables, based on what seemed conceptually relevant from the crafted profiles and by considering the variables and factors that were most common in our data and those with the highest variability across our sample. We used these factors on the IM residents' actual comments but unfortunately the results were disappointing – we found no clear correlations between the LIWC scores and the ITER numeric scores.[22] Similar findings arose from the DAL. However, using the software "out of the box" has its limitations – indeed, the LIWC dictionary is modifiable, and there may have been other variables that we excluded that may have been important. Further, it is possible that correlation with narrative comment scores would be higher than with numeric scores.

To take this research forward requires a deeper understanding of both the nature of the language we use in assessment and of the software programs. This is what led me to pursue a PhD, with the main research questions as stated above.

3b. Approach (method and setup)

In this section I will outline several research studies that will address these questions. However, as I've found with prior research programs, the logic, sequence and relevance of these projects may change considerably as the program progresses, depending on what is learned as data are collected. These studies together may best be thought of as not necessarily linear or sequential (indeed, some will run concurrently) but rather are intended as a triangulation of methodologies and approaches, each contributing a unique perspective.

First study
Goal – To better understand what faculty infer from reading comments about residents; i.e., how can we explain the high inter-rater reliability of faculty rank-ordering the residents based only on the comments? What did they "see" in the language, what did they think their colleagues were trying to convey about their residents, what competencies seemed most important in forming their opinions, etc.?

Method – Analysis will be conducted of 24 interviews conducted with faculty attendings as they rank-ordered and categorized narrative comments from ITERs. These interviews yielded 150 single-spaced pages and a preliminary read-through confirms that the transcripts are detailed and rich. Several possible methods for analysis will be explored, starting with a constructivist grounded theory approach[23] using sensitising concepts gleaned from recent studies.[11] Other approaches may involve framework analyses (e.g., using CanMEDs as a template to search for talk on competencies), or searches for specific terms in the language (e.g., based on variables from the LIWC), etc. Because the data are so detailed and rich this will take several months to complete and may result in more than one set of findings.

Outcome –One (or more) conceptual frameworks for understanding what faculty "read into" the comments about residents and how they "decode" the messages written by their colleagues. This will lead to better understanding of how IM faculty conceptualize competence and performance, especially in those domains not currently captured on ITERs. Further, it will lead to a better understanding of how to modify the LIWC/DAL analysis (e.g., by suggesting variables that may be important, etc).

Second set of studies
Goal – To extend findings from work at the University of Toronto in a number of ways by conducting similar methods at the University of British Columbia (UBC) and adding experimental manipulations to

the process. UBC was chosen as it also has a large IM program (~160 residents over 3 years, second only to Toronto) and their ITER forms are quite similar to those at Toronto. Of note are two key differences: the UBC ITERs have space for comments after each CanMEDs role as well as an overall comment box at the end; and the program is distributed, including rotations at several remote locations.

Method – The first protocol will be the same as the one used at Toronto. Full details can be found elsewhere[2] but briefly, faculty will be recruited to rank-order the comments extracted from residents' ITERs and will be interviewed regarding their decisions. The same process for analysis will be used as for the Toronto data, including a factor analysis of numeric data and assessment of reliability of the ITERs and of the comments. It will also be possible to create a predictive model of resident performance in PGY3 by using PGY1 and PGY2 ITER scores as well as comment scores. The interviews will be analyzed with the approach determined based on findings from Study 1.[23]

Related protocols will involve experimental manipulation of the comment documents in order to test several hypotheses. The first will focus on determining how much commentary needs to be read in order for attendings to form stable impressions. So far we have used aggregates of 8-10 ITERs' worth of comments per resident, but similar to numeric ratings our high inter-rater reliabilities may have arisen largely from the aggregation of the data. We will attempt to determine the number of comment boxes necessary to produce adequate reliability (e.g., a single rotation, 3, 6 or all) both in terms of faculty judgment and LIWC/DAL scores.

A second study will focus on language related to competencies. Because the UBC ITERs provide room for comments after each CanMEDs role it will be possible to assess, for example, whether a certain comment or adjective means the same thing if linked to one role compared to another. This study will use comments labelled as specific CanMEDs roles or presented in aggregate without labels and will use a similar rank-ordering protocol. Comparison data can be generated from the U of T ITERs by extracting data explicitly related to specific competencies and creating new narrative profiles. These sets of profiles (with and without competency labels) can be tested with a new sample of faculty from both Toronto and UBC.

Outcomes – One outcome will be either validation or modification of the larger theory that faculty attendings in IM have a shared understanding/conceptualization of performance of medical residents, this time across two distinct yet similar institutions. Secondarily we will have advanced our understanding of what makes narrative comment analysis so apparently reliable and how to optimize the efficacy and feasibility of this sort of analysis. Results will also inform the debate regarding the utility of treating competencies and roles as though they are "speciated".

Third set of studies

Goal – To explore the utility of the LIWC and or DAL as methods to analyze the language on assessments in a way that reflects meaning from the assessors' point of view.

Method – A dataset will be created consisting of the ITERs of two cohorts of IM residents at U of T (approx. 200 per cohort) and one from UBC (approx 160). The comments will be extracted and prepared for use in the LIWC and DAL programs. Experience indicates that this step is labour intensive but critical to the functioning of the programs. Because the two programs are quite different they will be tested separately. For the LIWC, in order to determine which factors/variables to include I will use findings from the qualitative interviews and other sources (e.g., studies and literature from other domains and consultation with the developers) to create a list of conceptually relevant factors. These will be supplemented with measures from the data such as relative prevalence and variability of factors across ITERs. For the DAL, analysis will involve the 3 basic dimensions described.

Several analytic approaches will be explored. The first will involve basic analyses and descriptive findings from each program. For the DAL I will explore the range of emotion, activation and concreteness of imagery in ITER comments, and will answer questions such as: Is language more active or concrete when describing excellent or struggling residents; is emotion more neutral or lukewarm in mediocre performers, etc.? For the LIWC, relevant analyses will explore the types of positive and negative emotion words present in the comments, how attendings use pronouns and other telling parts of speech in their comments, etc. For both programs I will then conduct analyses to determine whether any of these dimensions differ based on factors such as residents' numeric scores or rank-order within their programs, and will determine consistency of various language dimensions across an individual resident's rotations. I will also assess whether the correlations between LIWC and DAL factors and narrative comment rankings will be better than correlations with numeric scores, as described above.

Preliminary analyses suggest there are differences in some language dimensions between PGY1, 2 and 3 years, which may reflect attendings' conceptualizations of what is important at each stage. Therefore another approach will involve an attempt to identify distinct profiles, or pictures of residents, depending on stage of training and performance, that may be helpful in the creation of milestones or EPA's.

Finally, it may prove interesting to test and refine these approaches on similar types of data from other sources, such as teacher evaluations, as there is greater range and more variability in these scores and more negative comments in this type of data, at least in our experience.[24]

Outcome – An in-depth understanding of the potential utility of using language software to evaluate narrative comments. This could lead to better feedback, flagging of individuals with issues, identification of trends/patterns over time that may indicate developing weaknesses, etc. This will also contribute to our understanding of how faculty and others conceptualize competence, in a way they might not even be aware of (e.g., unconscious use of certain language, phrases, etc).

3c. Literature references

max. 35 references.

1. Chaudhry SI, Holmboe ES, Beasley BW. The state of evaluation in internal medicine residency. J Gen Intern Med. 2008; 23:1010-1015.

2. Ginsburg S, Eva KW, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. Acad Med. 2013; 10.

3. Hodges BD. Assessment in the post-psychometric era: Learning to love the subjective and collective. Med Teach. 2013; 35:564-568.

4. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. Med Educ. 2011; 45:560-569.

5. Crossley J, Jolly BC. Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. Med Educ. 2012; 46:28-37.

6. Hodges BD, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med. 1999; 74:1129-1134.

7. Sterkenburg A, Barach P, Kalkman C, Gielen M, ten Cate O. When do supervising physicians decide to entrust residents with unsupervised tasks? Acad Med. 2010; 85:1408-1417.

8. Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? Teach Learn Med. 1993; 5:10-15.

9. Durning SJ, Hanson J, Gilliland WM,J.M., Waechter D, Pangaro LN. Using qualitative data from a program director's evaluation form as an outcome measurement for medical school. Mil Med. 2010; 175:448-452.

10. Regehr G, Ginsburg S, Herold J, Hatala R, Eva KW, Oulanova O. Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance. Acad Med. 2012; :427.

11. Ginsburg S, Gold W, Cavalcanti R, Kurabi B, McDonald-Blumer H. Competencies "plus": The nature of written comments on internal medicine residents' evaluation forms. Acad Med. 2011; 86:s30-s34.

12. Frohna A, Stern DT. The nature of qualitative comments in evaluating professionalism. Med Educ. 2005; 39:763-768.

13. Canavan C, Holtman MC, Richmond M, Katsufrakis PJ. The quality of written comments on professional behaviors in a developmental multisource feedback program. Acad Med. 2010; 85:S106-9.

14. Kiefer CS, Colletti JE, Bellolio MF, et al. The "good" dean's letter. Acad Med. 2010; 85:1705-1708.

15. Kan Ma H, Min C, Neville A, Eva KW. How good is good? students and assessors' perceptions of qualitative markers of performance Teach Learn Med. 2013; 25:15-23.

16. Bogo M, Regehr C, Woodford M, Hughes J, Power R, Regehr G. Beyond competencies: Field instructors' descriptions of student performance. J Soc Work Educ. 2006; 42:579-593.

17. Ginsburg S, McIlroy J, Oulanova O, Eva KW, Regehr G. Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. Acad Med. 2010; 85:780-786.

18. Whissell C. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language Psychol Rep. 2009; 105:509-521.

19. Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ. The Development and Psychometric Properties of LIWC 2007 <http://liwc.net/liwcdescription.php>. Accessed Dec. 9. 2007.

20. Pennebaker JW, Francis M. Cognitive, emotional, and language processes in disclosure. Cognition & Emotion. 1996; 10:601-626.

21. Pennebaker JW. The Secret Life of Pronouns: What our Words Say about Us. 1st ed. New York, NY: Bloomsbury Press, 2011.

22. Ginsburg S, Eva KW, Regehr G. Exploring a Novel Method for Analyzing Written Comments on Residents' ITERs. 2012; .

23. Charmaz K. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. London: Sage Publications, 2008.

24. Ginsburg S, Brydges R, Imrie K, Lorens E. The Nature of Residents' Written Comments on Teaching Evaluation Forms. Med Educ. 2012; 46 (Supplement s1):61-61.

3d. Time Plan

max. 1 page.

As I already have peer-reviewed funding for part of this work I will be able to begin immediately.

First study – Analysis of the interview transcripts can begin immediately. Based on my experience it will take at least several months to fully analyze and report on themes. I hope to have a first draft of a paper written by early 2014.

Second set of studies – Ethics approval had been obtained but will need to be renewed prior to starting data collection. I will coordinate with my collaborators at UBC (Glenn Regehr and Kevin Eva) to hire a suitable research assistant. They will also assist me in obtaining the relevant database of evaluations. I will then travel to UBC to train the RA and begin data collection. I will conduct the first few experiments and interviews with the RA but given the complexities of recruitment especially across more remote sites I imagine the RA will collect most of the data. Another option is to conduct parts of the exercise by video-call (e.g., Skype) depending on each protocol. Given the start-up time required I anticipate having the ITER data compiled by November, 2013 and plan to start data collection from participants at UBC in the spring of 2014. Analysis will proceed concurrently with data collection and will continue afterwards, and will take a few months in total for the quantitative and qualitative portions.

Third set of studies – The in-depth exploration and analyses using the LIWC and DAL programs, along with external consultation with the authors, will run concurrently with the above and can start at any time. Preliminary analyses will be conducted with data already compiled at U of T (we have available the comments from PGY1-3 for two complete cohorts of residents) which will be supplemented with new data collected at both Toronto and UBC. I expect this phase to take 1-2 years given the complexities of the programs.

3e. Scientific setting

Main publications of applicant(s): (Please note: only relevant publications listed here)

1. **Ginsburg S**, Eva K, Regehr G. Do In-Training Evaluation Reports Deserve Their Bad Reputations? A Study of the Reliability and Predictive Ability of ITER Scores and Narrative Comments. Academic Medicine. In Press.
2. Stroud L, Oulanova O, Szecket N, **Ginsburg S**. "The benefits make up for whatever is lost": Residents' altruism and accountability in a new call system. Academic Medicine. 2012; 87(10):1421-1427.
3. **Ginsburg S**, Bernabeo E, Ross K, Holmboe E. It depends: Results of a Qualitative Study Investigating How Practicing Internists Approach Professional Dilemmas. Academic Medicine. 2012; 87(12):1685-1693.
4. Choo, EK, Fernandez R, Hayden E, Clyne B, Schneider J, **Ginsburg S**, Gruppen L. An Agenda for Increasing Grant Funding of EM Education Research. Academic Emergency Medicine. 2012;19(12):1434-1441..
5. **Ginsburg S**. Duty Hours Viewed Through a Professionalism Lens. BMC Medical Education. In Press..
6. Bernabeo, E.C., Holmboe, E.S., Ross, K., Chesluk, B., & **Ginsburg, S**. The Utility of Vignettes to Stimulate Reflection on Professionalism: Theory and Practice. Advances in Health Sciences Education. 2012; Online first, DOI 10.1007/s10459-012-9384-x.
7. Regehr G, **Ginsburg S**, McIlroy J, Hatala R, Eva K, Oulanova O. Using "Standardized Narratives" to Explore New Ways to Represent Faculty Opinions of Resident Performance. Academic Medicine. 2012;87(4):416-427.
8. Ho M-J, Lin C-W, Chiu Y-T, Lingard L, **Ginsburg S**. Professionalism in relations: a cross-cultural study of students' approach to professional dilemmas – sticks or ripples. Medical Education. 2012:46(3):245–256.
9. **Ginsburg S**. Respecting the expertise of clinician assessors: Construct-alignment is one good answer. Medical Education. 2011;45(6):546-548.
10. **Ginsburg S**, Gold W, Cavalcanti R, Kurabi B, McDonald-Blumer H. Competencies "Plus": The Nature of Written Comments on Internal Medicine Residents' Evaluation Forms. Academic Medicine. Acad Med 2011;86(10 suppl):s30-s34..

11. Mylopoulos M, Regehr G, **Ginsburg S**. Exploring residents' perceptions of expertise and expert development. Academic Medicine. 2011;10(Suppl):s46-s49..
12. Blissett S, Law C, Morra D, **Ginsburg S**. The relative influence of available resources during the residency match: A national survey of Canadian medical students. Journal of Graduate Medical Education. 2011;3(4):497-502.
13. Hodges B, **Ginsburg S**. Assessment for professionalism: Consensus statement and recommendations from the Ottawa 2010 conference. Medical Teacher. 2011;33(5):354-363.
14. **Ginsburg S**, Lingard L. "Is That Normal"?: Preclerkship Students' Approach to Professional Dilemmas. Medical Education. 2011;45(4):362-371.
15. Bernabeo E, Holtman M, Rosenbaum J, **Ginsburg S**, Holmboe E. Lost in Transition: A Study of the Hidden Curriculum of Transitions in Residency Education. Academic Medicine. 2011;86(5):591-598..
16. Holmboe E, **Ginsburg S**, Bernabeo E. Confronting Our Assumptions: The Rotational Approach to Medical Education. Medical Education. 2011;45(1):69-80.
17. Bryden P, **Ginsburg S**, Kurabi B, Ahmed N. Professing Professionalism: Are We Our Own Worst Enemy? Faculty's Experience of Teaching and Evaluating Professionalism in Medical Education. Academic Medicine. 2010;85(6):1025-34..
18. **Ginsburg S**, McIlroy J, Oulanova O, Eva K, Regehr G. Towards Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competencies. Academic Medicine. 2010;85(5): 780-786.
19. **Ginsburg S**, Regehr G, Mylopoulos M. From Behaviours to Attributions: Further Concerns Regarding the Evaluation of Professionalism. Medical Education. 2009;43(5):414-425.
20. Morra D, Regehr G, **Ginsburg S**. Medical Students, Money and Career Selection: Medical Students' Perceptions of Remuneration in Family Medicine. Family Medicine. 2009:42(2):105-110.
21. **Ginsburg S**, Lingard L, Regehr G, Underwood K. Know When to Rock the Boat: How Faculty Rationalize Students' Behaviours. Journal of General Internal Medicine. 2008:23(7):942-947..
22. Morra D, Regehr G, **Ginsburg S**. Anticipated Debt and Financial Stress in Medical Students. Medical Teacher. 2008:30(3):313-315 (Trainee publication, Dante Morra).
23. **Ginsburg S**, Regehr G, Mylopoulos M. Reasoning When it Counts: Students' Rationales for Action on a Professionalism Exam. Academic Medicine. 2007;82(10 Suppl):S40-43**.**
24. Cruess R, McIlroy J, Cruess S, **Ginsburg S**, Steinert Y. The P-MEX (Professionalism Mini Evaluation Exercise): A Preliminary Investigation. Academic Medicine. 2006;81(10 suppl):S74-78.
25. Arnold L, Shue CK, KRitt B, **Ginsburg S**, Stern DT. Medical Students' Views on Peer Assessment of Professionalism. Journal of General Internal Medicine. 2005:20(9):819-24 (See also Response to Letter to the Editor, J Gen Intern Med 2006:21(4):399).
26. **Ginsburg S**, Kachan N, Lingard L. Before the White Coat: Perceptions of Professionalism in the Pre-clerkship. Medical Education. 2005:39(1): 12-19.
27. **Ginsburg S**, Stern DT. The Professionalism Movement: Behaviours are the Key to Progress. American Journal of Bioethics. 2004;4(2):14-15.
28. **Ginsburg S**, Schreiber M, Regehr G. The Lore of Admissions Policies: Contrasting Formal and Informal Understandings of the Residency Selection Process. Advances in Health Sciences Education. 2004;9:137-145..
29. Lavine, E, Regehr, G, **Ginsburg, S**. The Role of Attribution to Clerk Factors and Contextual Factors in Supervisors' Perceptions of Clerks' Behaviours. Teaching and Learning in Medicine. 2004;16(4):317-22. **Ginsburg S**, Regehr G, Lingard L**.** Basing the Evaluation of Professionalism on Observable Behaviours: A Cautionary Tale. Academic Medicine. 2004; 79(10 suppl): S1-4.
30. **Ginsburg S**, Regehr G, Lingard, L. To Be and Not To Be: The Paradox of the Emerging Professional Stance. Medical Education. 2003;37(4):350-357.
31. **Ginsburg, S**, Lingard, L, Regehr, G. The Disavowed Curriculum: What Motivates Students to Act in Professionally Challenging Situations. Journal of General Internal Medicine. 2003;18(12):1015-22.
32. **Ginsburg S,** Regehr G, Stern DT, Lingard, L. The Anatomy of the Professional Lapse: Bridging the Gap between Traditional Frameworks and Students' Perceptions. Academic Medicine. 2002;77(6):516-522.
33. **Ginsburg S**, Regehr G, Hatala R, McNaughton, N, Frohna, A, Hodges, B, Lingard, L, Stern, D. Context, Conflict, and Resolution: A New Conceptual Framework for Evaluating Professionalism. Academic Medicine. 2000; 75(10 suppl): S6-11.

Relevant Published Abstracts

1. **Ginsburg S**, Brydges R, Imrie K, Lorens E. The Nature of Residents' Written Comments on Teaching Evaluation Forms. Medical Education. 2012;46(Supplement s1):61..
2. Brydges R, **Ginsburg S**, Imrie K, Lorens E. How Should We Change our Teacher Evaluation Forms? Lessons from One Department of Medicine. Medical Education. 2012;46(Supplement s1):61

Relevant Books and Chapters.

1. Levinson W, **Ginsburg S**, Hafferty F, Lucey C. Understanding Professionalism in Medicine (working title). Commissioned by Lange (McGraw Hill); expected publication date of April, 2014.
2. **Ginsburg S**. Chapter 41. Assessing Professionalism in Medical Education. In A Practical Guide for Medical Teachers, 4th edition (Eds. Dent J, Harden R).2013
3. Hodges, B; **Ginsburg, S**. Assessment of Professionalism. International Best Practices for Evaluation in the Health Professions. In Press.
4. **Ginsburg S**, Lingard L. Using Reflection and Rhetoric to Understand Professional Behaviours. For "Measuring Medical Professionalism". Stern DT, Ed. 2006 Oxford University Press NY, NY.

3f. Setting within Research Group
(other relevant research, proposal part of a research program)
max. 1 page.

I have already included elsewhere in this proposal relevant findings from related and foundational work so do not wish to be redundant. My publications are all listed above, in 3^{e}, as requested. However, it may be worth noting that I have extensive research experience, having previously led a successful research program focused on understanding and evaluating professionalism. This program has an excellent track record of obtaining peer-reviewed funding and has led to a few dozen articles and several research awards. To achieve this I have always worked with a team and have developed strong collaborative relationships with several colleagues, some of whom will be involved with various studies proposed here.

3g. Output
Expected scientific output and dissemination of results
(Thesis, papers, presentations, dissemination plan).

The first study will lead to at least one qualitative paper suitable for publication in a peer-reviewed journal.

The second set of studies will lead to several papers including one on the results of the rank-ordering exercise and interviews at UBC, one on the determination of the amount of commentary required to form stable opinions and one on the analysis of competency-related vs. aggregate comments.

The third set of studies will lead to papers involving the findings from the LIWC and DAL analyses on the ITER (and potentially teacher evaluation) data from both UBC and Toronto.

I am also considering an additional review-type paper discussing the analysis of language in assessment from a theoretical and empiric perspective.

In addition to the proposed papers I anticipate submitting abstracts to relevant scientific meetings as results begin to emerge, such as Research in Medical Education at the AAMC Annual Meeting, the Canadian Conference on Medical Education, and perhaps AMEE and/or the Ottawa conferences as well. I am scheduled to participate in a pre-conference workshop at the forthcoming Ottawa conference (April, 2014) on the topic of "Construct alignment in judgment-based assessment".

3h. Societal & Scientific Relevance
(if applicable)
max. 1 page.
How can results be applied in other research areas?
How can results be applied in society, business, etc.?

In parallel work I have been exploring the narrative comments on teacher evaluation forms and I believe that whatever is learned during this proposed research will be quite applicable to that setting. There are many similarities, including similar forms containing both numbers and comments, based on a block of exposure (2-4 weeks), and being heavily criticized for being too subjective. I have already analyzed over 1000 forms using NVivo software (its word count and textual analysis features) and presented results at a national conference, but have yet to publish this work. Depending on the success (or failure) of using the LIWC and DAL I would plan on using similar protocols to analyze teacher data in the future.

Signature

Name: Shiphra Ginsburg

Place: Toronto, Ontario, Canada

Date: September 4th, 2013