The Development of an Online Practice-Based Evaluation Tool for Social Work

Regehr, Cheryl; Bogo, Marion; Regehr, Glenn

Version  Post Print/ Accepted Manuscript


How to cite TSpace items

Always cite the published version, so the author(s) will receive recognition through services that track citation counts, e.g. Scopus. If you need to cite the page number of the TSpace version (original manuscript or accepted manuscript) because you cannot access the published version, then cite the TSpace version in addition to the published version using the permanent URI (handle) found on the record page.
The Development of an Online Practice-Based Evaluation Tool for Social Work

Cheryl Regehr (Corresponding Author)
Vice-Provost, Academic Programs
Professor, Factor-Inwentash Faculty of Social Work,
University of Toronto
27 King’s College Circle
Toronto, M5A 1S1
Cheryl.regehr@utoronto.ca

Marion Bogo
Professor, Factor-Inwentash Faculty of Social Work,
University of Toronto
Marion.bogo@utoronto.ca

Glenn Regehr
Professor, Faculty of Medicine
University of British Columbia
Glenn.regehr@ubc.ca

This research was generously supported by a grant from the Social Sciences and Humanities Research Council of Canada
The Development of an Online Practice-Based Evaluation Tool for Social Work

Abstract

Objective: This paper describes the development of a practice-based evaluation tool that allows instructors to represent their student’s clinical performance in a way that is sufficiently authentic to resonate with both instructors and students, is psychometrically sound and is feasible in the context of real practice.

Method: A new online evaluation tool was designed to address several of the problems associated with previous methods of evaluation and was tested on 190 field instructor-student pairs.

Results: Results demonstrated feasibility of the tool, high acceptability from students and faculty, high internal consistency and clearly reduced ceiling effect when compared with a traditional competency-based evaluation (CBE) tool. It did however continue to result in a strong skew towards positive evaluation and did not increase the identification of students at risk.

Conclusions: The online practice-based evaluation tool demonstrates promise in redressing some of the evaluation issues posed by the previous CBE model of evaluation.

Key words: competency, practice evaluation, field education, clinical training
The Development of an Online Practice-Based Evaluation Tool for Social Work

As with all disciplines within the academy, social work educators carry the responsibility of imparting knowledge and encouraging critical and creative thinking in their students. Consistent with other professions we also carry the responsibility for teaching the relevant skills and technical knowledge of the profession. In addition however, we carry the responsibility for ensuring that graduates entering the profession will not pose a risk to the members of the public who become their clients. The assessment of student competence is thus of paramount importance. Yet the complexity of this endeavor poses a challenge (Bogo, et al, 2004; Regehr et al, 2007). When is a student at a point where he or she can be permitted to function as an independent practitioner? How do we reliably measure competence? What indicators of competence as measured in the academy are valid predictors of competent performance once a student enters independent practice?

Attempts to develop reliable and valid instruments and processes in social work and in other professions have been hampered by the lack of standardized definitions of competency and the lack of agreed upon performance indicators (Bondy, et al., 1997; Sliwa & Kowalske, 2000; Siegel & Greenberg, 2000). Poor inter-rater reliability on assessments of student performance have been documented (Bogo, et al, 2002; Gonsalvez & Freestone, 2007; Govaerts, et al, 2006). There is consistent research that suggests that assessors regularly give above average ratings resulting in a ceiling effect, despite obvious differences in performance among the students (Bogo, et al, 2002;

Many of the current measurement tools for evaluating performance emanate from a Competency Based Education (CBE) perspective in which skills are identified in clear measurable terms (Arkava & Brennan, 1976; Brennan, 1982). Despite the appeal of this model, critics have suggested that reducing practice to discrete skills loses the very essence of professionalism (Hyland, 1995; Williams, 1997). Rather, competence rests upon integrated knowledge, frequently referred to as highly developed intuition, which incorporates critical reflection, practice experience, discipline specific theories, and empirical research (Bogo & Vayda, 1998).

In a recent program of research in social work, the authors began the process of developing authentic clinical assessment tools that reflected the language and constructions of competence that field instructors in social work used when describing high performing, typical, and problematic students in their own words. The first effort in this program involved the development of a six-dimension, 5-point rating scale in which the dimensions were derived from a grounded theory analysis of faculty descriptions of real students they had supervised. The anchors for the scale were drawn from the actual language faculty used in their descriptions of students performing at different levels (Bogo et al, 2004; Regehr, et al, 2007). This tool was called the Practice-Based Evaluation (PBE) tool.
The clear failure of this new instructor-oriented practice-based scale to demonstrate different evaluation properties (Regehr et al, 2007) compared to the traditional scale competency-based evaluation (CBE) used by the Faculty at the time (whose properties were previously described in Bogo et al, 2002), led to a second effort. This second effort involved 20 brief (three paragraph) vignettes representing iconic descriptions of students, that had previously been ranked according to level of performance by experienced field instructors. A subsequent group of field instructors were asked to identify the vignette(s) that most closely resembled their most recent student. An individual student score was then generated based on the level of performance represented by the vignettes selected. While this evaluation system was more successful in spreading out student scores relative to the traditional competency-based evaluation tool and new practice-based scale described above (Regehr et al, 2007), there was doubt expressed among instructors regarding the feasibility of the “vignette matching” process as a regular evaluation activity.

The current paper, therefore, describes the third effort to evolve a practice evaluation tool that allows instructors to represent their student’s clinical performance in a way that is sufficiently authentic to resonate with both instructors and students, is psychometrically sound and is feasible in the context of real practice. This new online evaluation tool was intended to deal with several of the problems associated with previous methods of evaluation. 1) It seeks to reduce the ceiling effect and create greater disbursement of performance scores by eliminating the explicit use of numerical grading; 2) It seeks to more reliably identify students whose level of competence may
represent a risk to the public; 3) It seeks to increase the practice validity of measurement of performance by using phrases generated by clinical instructors.

**Methods**

**Tool Development**

The new description-based evaluation tool was developed based on the details of the original 6-dimension Practice-Based Evaluation (PBE) tool, the construction of which was described in (Regehr, 2007). Briefly, to construct a scale that reflected instructors’ views of competence, in-depth interviews were held with 19 experienced field instructors who were asked to provide detailed descriptions of an exemplary, an average, and a problematic student that they had supervised in the field practicum. Grounded theory analysis of the 57 student descriptions revealed six instructor-relevant constructs consistently mentioned in the interviews when describing all levels of student performance. These six instructor-identified constructs were selected as the six dimensions of practice competence to be evaluated with the tool. The resulting dimensions were: learning and growth, behavior in the organization, clinical relationships, conceptualizing practice, assessment and intervention, and professional communication.

The next task was to anchor each dimension with a set of detailed statements describing student behavior and skills at different levels of performance. In order to do so, 20 iconic vignettes were developed from field instructor descriptions of their students using verbatim language drawn from the instructor interviews. These vignettes were then grouped into five performance levels and ranked from one to 20
based on levels of competency by 10 additional field instructors. This process resulted in an inter-rater reliability of 0.83 (Bogo, et al, 2004). In the previous study this ranking and sorting process was accomplished by the instructors with an inter-rater reliability of 0.83 (Bogo, et al, 2004). Statements from the ranked and categorized vignettes served as indicators at the five levels of performance in each of the six dimensions.

The new online tool was developed to capitalize on the strengths of the PBE tool but attempted to address its shortcomings. Of key importance, in our progressive studies on evaluation, we became convinced that the use of numerical ratings continued to result in ceiling effects – that is regardless of the type of scale used, the majority of students were rated as being at the very highest levels of performance. Thus the new scale eliminated numerical ratings from the view of the evaluator. The phrases that were acting as anchors at all five levels of performance for a given dimension were collated into a single list and placed in alphabetical order with no indication of a numeric value in order to avoid implying an ordinal quality to the dimensions.

**Evaluation Process**

For each dimension, the evaluator (instructor and student) was required to select from the list of descriptive phrases those that best described the student using a pull down menu (see Figure 1). To create an overall score for the student, once all relevant phrases were selected for each dimension, and the evaluation was submitted, a numerical value of 1-5 was assigned to each selected phrase based on level of performance at which the phrase was an anchor in the original PBE tool. A mean score
was derived for each dimension and a final score was calculated as the unweighted average of the 6 dimensional scores.

Students and instructors both complete the form independently and then review the ratings of the other only after they have submitted their own assessment online (see Figure 2). Instructors are thus not pressured to change the scores based on a negotiation process with the student.

Procedure

All aspects of the study were approved by the Health Sciences Research Ethics Committee of the University of Toronto prior to implementation.

In the spring term of 2008, a pilot study was implemented for 20 students in three affiliated teaching hospitals with all students and field instructors consenting to participate. Feasibility was assessed, and participants provided feedback on their views of the tool. Some minor variations were required with respect to the online interface. Overall however, feedback was positive.

Starting in the fall term of 2008, the web-based descriptive PBE tool was established as the formal field placement evaluation tool for all second year MSW students in the Faculty (n=190). The first institutional use of the tool occurred at the midterm of the student placements at the end of December 2008. The second full use of the tool occurred at the end of placement for all students at the end of April 2009.

An information form was placed at the beginning of the evaluation form and both field instructors and students were asked whether their data could be used to assess the utility of the new tool as part of a research study. If a student or instructor
chose not to participate, their data was removed from the research version of the
database by a research team member who was in no way connected to the Faculty and
therefore had no knowledge of, or influence over, the future success of the students or
field instructors. All evaluations, regardless of consent to participate in the research
aspect of the process, were used as the formal evaluation of all students.

Finally, a series of information and feedback sessions were held with field
instructors and students to obtain their views of the new tool and process for
evaluation.

Results

Of 780 potential evaluations (190 students and 190 instructors at each of two
times), a total of 611 evaluations were completed and consented. These broke down
into 151 completed evaluations from field instructors and 157 complete evaluations
from students at the midterm and 152 completed evaluations from field instructors and
151 complete evaluations from students at the time of the final evaluation.

Reliability of the New Measure

To assess the internal consistency of the six subscale scores, an “average-score”
 intra-class correlation coefficient was calculated using Shrout and Fleiss’ (1979) Case 1
model. To assess the reliability of the total scale score across the six subscales, the mean
value of each subscale was generated for each student, and Cronbach’s alpha was
calculated for the total score across the six scales. Table 1 presents these reliabilities
separately for the instructors and the students at both the midterm (table 1a) and final
evaluation (table 1b) time points. These results suggest that, both within and across the
subscales, the statements selected by both field instructors and students represent similar levels of performance. We note with interest that, given that actual numbers were not visible to the raters when making their selections, these results would appear to provide some evidence that these dimensions of performance actually are correlated with each other rather than simply being indistinguishable due to rater tendencies to select similar numerical ratings across indicators.

### Table 1a: Reliability of the New Measure at Midterm Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Learning and Growth</th>
<th>Behavior in the organization</th>
<th>Clinical relationships</th>
<th>Conceptualizing practice</th>
<th>Assessment and intervention</th>
<th>Professional communication</th>
<th>6-dimension alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Instructors (n = 151)</strong></td>
<td>0.662</td>
<td>0.532</td>
<td>0.594</td>
<td>0.466</td>
<td>0.697</td>
<td>0.669</td>
<td>0.907</td>
</tr>
<tr>
<td><strong>Students (n = 157)</strong></td>
<td>0.545</td>
<td>0.428</td>
<td>0.596</td>
<td>0.407</td>
<td>0.610</td>
<td>0.655</td>
<td>0.934</td>
</tr>
</tbody>
</table>

### Table 1b: Reliability of the New Measure at Final Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Learning and Growth</th>
<th>Behavior in the organization</th>
<th>Clinical relationships</th>
<th>Conceptualizing practice</th>
<th>Assessment and intervention</th>
<th>Professional communication</th>
<th>6-dimension alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Instructors (n = 152)</strong></td>
<td>0.587</td>
<td>0.521</td>
<td>0.663</td>
<td>0.516</td>
<td>0.642</td>
<td>0.690</td>
<td>0.997</td>
</tr>
<tr>
<td><strong>Students (n = 151)</strong></td>
<td>0.579</td>
<td>0.400</td>
<td>0.645</td>
<td>0.479</td>
<td>0.582</td>
<td>0.664</td>
<td>0.978</td>
</tr>
</tbody>
</table>

**Comparisons from Midterm and Final Evaluation Scores**

Of all the evaluations completed by faculty at the two time points, 134 faculty rated students at both midterm and final evaluation. The Pearson’s correlation between the faculty ratings at these two points was 0.45. This relationship is represented graphically in Figure 3. The diagonal line represents the x=y line, so data points above
the line represent students whose rating improved from the midterm to the final
evaluation and data points below the line represent students whose faculty rated them
lower at the final evaluation.

Comparisons Across Scales

We were able to compare the distribution of the 152 final evaluations generated
by the instructors using the new PBE scale in 2008/09 with the scores generated by
instructors using the old CBE scale on 300 second year students from 1992 to 1998.
There was a large positive skew in data using the CBE tool with fully 58% of the ratings
falling between 4.75 and 5.00 on the scale. By contrast, while there is still a skew in the
new data set using the Online PBE tool, this skew is substantially reduced, with the
modal response being the bin between 4.25-4.50 and with 8.5% of the scores being
below 4.00 in the new data set (compared to 4% in the old data set). A comparison of
the means represented by these data demonstrate that there is a significantly lower
mean score using the new scale relative to the old scale (see Table 2) with the difference
in means representing an effect size of approximately 0.76 standard deviations.

<table>
<thead>
<tr>
<th></th>
<th>CBE Tool</th>
<th>Online PBE Tool</th>
<th>Difference</th>
<th>t-value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4.70</td>
<td>4.46</td>
<td>0.24</td>
<td>7.67</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Std Dev</td>
<td>0.32</td>
<td>0.29</td>
<td>0.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>300</td>
<td>152</td>
<td>452</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparing Instructor and Student (self) Ratings
The data from these evaluations offers an interesting opportunity to explore student self-assessments in a novel way that would appear to circumvent some of the problems of self-assessment methodologies identified by Ward et al (2002). That is, since each student and each evaluator had the chance to select any set of descriptors from the 149 options, it is possible, for each individual student, to establish the concordance in these selections (the extent to which students selected the same descriptors that the faculty selected). This method is similar to “relative ranking” method of quantifying self-assessment performance described by (Regehr, et al, 1996). However because of the dichotomous nature of each item that could be selected or not, the appropriate statistic for measuring concordance of selections in this data set is Cohen’s kappa. Thus, for the 132 students who had both a final instructor evaluation and self-evaluation, a Cohen’s kappa was calculated. The distribution of these values is presented graphically in Figure 4 with a mean kappa score of 0.589. This suggests that although there was variance in the level of concordance across the student group, overall there was remarkably high concordance between individual students and their instructors regarding the descriptors selected to describe the individual.

Student and Field Instructor Response to the Tool

In an earlier paper (Bogo, et al, 2007) we described the results of focus group discussions in which field instructors identified challenges in providing corrective feedback to students. Instructors reported feeling a conflict between being supportive and empowering on one hand, and holding the responsibility for gate keeping on the
other hand. In addition however, they noted that negotiating ratings with the student in the process of evaluating practicum performance inevitably resulted in higher ratings.

Feedback sessions held related to the online PBE tool were overwhelmingly positive. Field instructors valued three aspects: 1) the ease with which the tool could be completed; 2) the fact that students and field instructors were forced to rate student performance independently; 3) the perceived authenticity of the statements. Students similarly valued the simplicity of the tool. However, a few students indicated that they were concerned that they did not know the numeric values of the scale when selecting items as they felt that would influence their selection. A concern expressed by both field instructors and students was that the PBE tool focused on clinical practice and was not applicable for macro social work practice. This is an issue that the researchers are currently addressing.

**Discussion**

The development of reliable and valid measures for assessing the competence of students is an ongoing challenge in social work despite the widely recognized need for such instruments to ensure the safe provision of services to the clients we serve. Previous attempts of developing measures have been plagued by problems that include poor reliability and lack of differentiation between students at different levels of competence caused in part by numerical rating scales that are confounded by relational issues between the supervisor and the student and then tendency towards leniency. This paper describes the development of a new online tool that eliminates the use of a
numerical system in the rankings, uses language derived from supervisors themselves and removes negotiation over evaluations through a separation of the process.

The results of this study demonstrate the overall success of the new model. Qualitative feedback from field instructors and students both during the pilot stage and during the implementation of the new tool resulted in positive responses both on the user friendly nature of the tool and the usefulness in providing student feedback. Further, the new tool was demonstrated to have good psychometric properties. There was high internal consistency across the various dimensions of the tool. That is, both within and across subscales the statements selected by both field instructors and students represent similar levels of performance. In addition, there was differentiation between student scores at mid-term and the final evaluation, indicating the ability of the tool to identify changes in performance as time elapsed and tasks and challenges within the placement changed.

The new tool was compared with a competency based tool that had been used by within the Faculty over the past 17 years. While the new tool continued to have a positive skew, the mean scores on the new tool were significantly lower and a substantially lower number of students were rated in the top category. Thus, the new tool clearly reduced the ceiling effect. The number of the students at the bottom end of the scale did not increase significantly. We are unsure if this is in part due to a sampling bias as both field instructors and students were required to consent to have data included in the analysis. It is possible that those who were struggling and thus possibly had lower scores, did not wish to have their data included. However, while we would
hope that few students have competency levels that would suggest that they are unsuited for the profession, a limitation of the new tool may continue to be that it does not identify those at risk.

The concordance between student and field instructor ratings was evaluated. Although there was high concordance between student and field instructor ratings, there was variation. Such variation is to be expected as previous research has demonstrated that while students and instructors may be in agreement about skill levels with respect to technical issues such as communication skills, individual and group treatment skills and community knowledge considerable disagreement can be found when there are problems with higher order or less concrete skills. These skills which are perhaps more reflective of personality styles include dealing with resistance, ambivalence or conflict in the therapeutic relationship, ability to function on a team and adherence to ethical standards (Regehr, et al, 2002).

In summary, the new online practice-based evaluation tool appears to have considerable promise for social work education. By hiding numerical rating scales from the evaluators, using language of expert field educators and having students and instructors complete the instruments independently, we have effectively reduced ceiling effects and increased the dispersion of scores across students in the tested cohort. This provides an exciting alternative to traditional CBE tools that bares replication in other social work programs. What remains unclear however, is the degree to which field instructor ratings of student performance and student self-ratings on the tool accurately predict professional performance as measured by independent raters, in terms of client
outcomes, or as a person develops experience in the field. The validity of the tool as a predictor of each of these outcomes requires further research and consideration.
References


Figure 1: New Online PBE Tool Evaluation Screen
Figure 2: New Online PBE Tool Report Screen
Figure 3. Scatterplot representing the correlation of instructor scores generated for each student at the midterm and final evaluation timepoints.
Figure 4. Frequency distribution of the Cohen’s kappa scores for each student as a representation of the concordance of student and faculty descriptor selections (as a measure of self-assessment performance for each student).
Fig 2
Fig 3