Using Self-Assessments to Detect Workshop Success

Do They Work?

Marcel D’Eon
University of Saskatchewan

Leslie Sadownik
University of British Columbia

Alexandra Harrison
Calgary Health Region

Jill Nation
University of Calgary

Abstract: An accepted gold standard for measuring change in participant behavior is third-party observation. This method is highly resource intensive, and many small-scale evaluations may not be in a position to use this approach. This study was designed to assess the validity and reliability of aggregated group self-assessments as one way to measure workshop effectiveness. In this study, participants completed a pre-, post-, and retrospective self-assessment on their perceived skill level in delivering feedback. Trained raters scored recorded role-play episodes. A statistically and practically significant difference in feedback skills was detected in both the self-assessments and observer ratings. The instruments used to assess participants’ feedback skills had acceptable reliability. Those charged with workshop evaluation should have some confidence that aggregated self-assessments can be used to help determine workshop effectiveness.

Keywords: aggregated self-assessments; workshop evaluation; skill assessment

One of the most challenging and resource-intensive tasks of trainers, developers, and teachers is the rigorous evaluation of educational interventions. For this reason, evaluations for most small-scale workshops are either not done or are completed superficially (Bamberger, Rugh, Church, & Fort, 2004; Brooks & Gersh, 1998; McEvoy & Buller, 1990; Pratt, McGuigan, & Katzev, 2000; Reid, Stritter, & Arndt, 1997; Rennie, 1996; Stone et al., 2003). Self-assessments, conversely, are relatively easy to obtain, require minimal resources, and can be used in the absence of baseline data (Bamberger et al., 2004; Pratt et al., 2000).

Marcel D’Eon, Director, Educational Support and Development, University of Saskatchewan, B103 Health Sciences Building, 107 Wiggins Road, Saskatoon, Saskatchewan S7N 5E5 Canada; (306) 966-2756; fax: (306) 966-7290; e-mail: marcel.deon@usask.ca.

Authors’ Note: Ethical approval from the University of Calgary was received for this study by Dr. C. Doig, Chair, Conjoint Health Research Ethics Board of the Faculties of Medicine, Nursing, and Kinesiology, University of Calgary and Affiliated Teaching Institutions. The Royal College of Physicians and Surgeons of Canada provided funding to conduct this study.

American Journal of Evaluation, Vol. XX No. X, Month XXXX xx-xx
DOI: 10.1177/1098214007312630
© 2008 American Evaluation Association
This study was designed to help establish the construct validity (Nachmias & Nachmias, 1987) of self-assessments for evaluating educational interventions so that trainers, developers, and evaluators may have more confidence in them.

For a variety of reasons (see Eva, Cunnington, Reiter, Keane, & Norman, 2004; Evans, McKenna, & Oliver, 2002; Kruger & Dunning, 1999; Orsmond, Merry, & Reiling, 1997; Ward, Gruppen, & Regehr, 2002) individual self-assessments, though considered an important skill for professionals (Evans et al., 2002; Fitzgerald, Gruppen, & White, 2000; Fitzgerald, White, & Gruppen, 2003; Orsmond et al., 1997) lack validity when used for summative and formative purposes (Albanese et al., 2006; Bardella, Janosky, Elicki, Ploof, & Kolarik, 2005; Barnsley et al., 2004; Davis et al., 2006; Evans, Leeson, Newton John, & Petrie, 2005; Vnuk, Owen, & Plummer, 2006; Wayne et al., 2006).

However intriguing this debate about the use of self-assessments for individual assessment purposes might be, this article does not argue at all for the use of self-assessments for this reason. The study reported here was based on aggregated self-assessments of individuals in the group and analysis of change in the group, not in individuals, and was used for workshop evaluation purposes. It is grouped self-assessments that may hold some promise to provide a valid and reliable indication of workshop effectiveness making the task of evaluating workshops and other educational sessions much easier and less expensive without sacrificing rigor, an unfortunate trade-off (Bernthal, 1995).

Other studies have tried to validate the use of self-assessment data to gauge workshop effectiveness (Howard, Millham, Slaten, & O’Donnel, 1981; Pratt et al., 2000; Skeff, Stratos, & Bergen, 1992) with promising results. Furthermore, many of the studies that have demonstrated no or low relationship between individual self-assessments and objective measures of learning or performance have failed to note that their grouped aggregated self-assessments and objective ratings were generally close (Albanese et al., 2006; Bardella et al., 2005; Evans et al., 2005; Fitzgerald et al., 2000; Vnuk et al., 2006; Wayne et al., 2006). As well, D’Eon (2004) demonstrated that trained raters had identified statistically significant improvements in teaching skills among postgraduate trainees (medical residents) after participating in a 2-day teaching workshop. These improvements were also identified in his earlier study (D’Eon, 1997) using retrospective self-assessments. This lends some evidence to the claim that grouped self-assessments might legitimately be used to evaluate workshop effectiveness.

Nevertheless, although there is potential in using the comparison of grouped pre- and postintervention self-assessments, many participants cannot assess themselves accurately, especially before the workshop (Eva et al., 2004; Pratt et al., 2000). To address this weakness, retrospective postintervention self-assessments have been used. These are obtained after the intervention and ask participants to “look back” and assess their level of knowledge and performance prior to the intervention. Skeff et al. (1992), Pratt et al. (2000), and Howard et al. (1981) found that the retrospective self-assessments correlated more closely with more objective ratings. This difference between the retrospective and preintervention self-assessments is known as the response-shift bias (Howard, 1980) and has been verified by Hyman (1993).

Although individual self-assessment data is fraught with controversy, some research has indicated that grouped self-assessments may provide a valid and reliable indication of workshop effectiveness. This study has therefore attempted to provide additional evidence that the difference between grouped retrospective and postintervention self-assessment scores (a) yields a reliable and valid measure of group change and, hence, (b) can be used as an indicator of intervention success (or failure).
Method

Context for the Study

Based on a needs assessment administered to about 20 medical residents at one institution, an educational workshop on delivering effective feedback was developed. The workshop was piloted at one and delivered at two other medical schools in Canada between October 2002 and September 2003. A total of four workshops were delivered in hospital settings to 61 medical residents from 2nd to 5th year across five different medical specialties. There was no specific exclusion criteria applied to participants; 10 residents did not attend both half-days of the workshop, and 5 others declined to participate in the study. The measured “outcome” of the workshop was their change in feedback skills immediately following the workshop. To establish construct validity, the results obtained through self-assessments were compared with the results obtained through trained third-party observers.

Data Collection

A criteria-based evaluation form outlining the observable characteristics of an effective feedback scenario was developed. The evaluation form contained 22 items that were grouped into four sections based on the framework for delivering effective feedback taught during the workshop with two summative items (overall rating and skill rating). The rating scale ranged from 1 (strongly disagree) to 6 (strongly agree). When the feedback skill was not observed, an “X” was entered into the score forms.

Immediately prior to the workshop, each medical resident participated in a role-play that required the resident to provide feedback to a “standardized student” played by an actor. (All actors had received training prior to the workshop.) This preworkshop encounter was videotaped. The preperformance was then self-assessed by each resident based on the standard scoring form.

After the initial role-play, residents participated in an interactive workshop consisting of 2 half-days. Immediately following the workshop, residents again participated in a different role-play feedback encounter with an actor. This encounter was also videotaped. Each resident then, from memory, assessed his or her own postworkshop performance using the same scoring form and was also asked to think back and “retrospectively” self-assess his or her first role-play performance. None of the residents viewed their videotaping either as part of the workshop or the study. Both sets of videotapes (pre- and postworkshop) were coded and then forwarded to two trained observers (the only people to view the videotapes), who then assessed the encounters independently using the same standardized scoring form also used by the residents. Because of technical difficulties, 4 sets of tapes were excluded, leaving a total of 42 complete sets of pre- and postworkshop recordings. The two observers were blinded to which recordings were made before or after the workshop. Each observer rated each recording.

Data Analysis

A paired sample t test was applied to the pre- and post-self-assessments and the observers’ ratings to identify a difference in resident feedback behavior before and after the workshop. Effect sizes were calculated. Correlation coefficients between the retrospective, pre-, and postworkshop self-assessments by residents, and observer ratings were determined. Missing data were handled with the pairwise deletion. The quality of the data was estimated using Cronbach’s alpha, a measure of internal consistency.
Forty-two complete data sets were obtained: one trainee’s pre- and post-self-assessment, the retrospective self-assessment, and the observers’ scoring sheets for both the pre- and postvideotaped role-plays for each individual trainee. The scores of the two trained raters were highly correlated (Pearson correlation $r = .96$, $p < .01$). We therefore averaged the scores of both observers to create one score for each recording. Cronbach’s alpha for pre- and postworkshop were as follows: for residents pre and post, .59 and .86, respectively; for external observers pre and post, .75 and .87, respectively. Cronbach’s alpha for residents’ retrospective self-assessment was .84. Both the observers’ and the residents’ scores resulted in moderate to high measures of internal consistency. The data (all except the preworkshop self-assessments) were of sufficient quality to be subjected to further analysis (Gall, Gall, & Borg, 2003). The low reliability of the preworkshop self-assessments does not necessarily weaken this study because these were not needed. The central analysis compared the grouped retrospective data (not preworkshop data) with the postworkshop self-assessments along with the ratings of the trained observers. More will be said about this in the Discussion section.

### Results

The mean scores (with 95% confidence intervals and standard deviations) are displayed in Table 1. Both groups scored the postworkshop performance significantly higher than the pre-workshop performance: residents $t = 7.08$, $p < .01$; external raters $t = 3.29$, $p < .0001$. The mean self-assessment scores of the residents and mean rating scores from the observers were not significantly different. The residents’ retrospective self-assessment scores were also not significantly different than their preworkshop self-assessment scores.

Using the residents’ retrospective and postworkshop self-assessment scores, the effect size was calculated to be .95. This is considered to be an indication of a large practical significance (Hojat & Xu, 2004). Using the scores from the trained raters, the effect size was less at .73. This falls just below what is thought to be a large effect size of crucial practical importance (.80) and well above the effect size considered to be of medium or of moderate practical importance (.50).

The Pearson correlations between the residents’ retrospective self-assessment with their preworkshop self-assessment was .54 ($p < .01$) and the correlation between the residents’ retrospective self-assessments and the trained raters’ preworkshop observations was $-.10$.

### Discussion

This study was designed to examine the construct validity of self-assessments for evaluating the effectiveness of educational interventions such as short workshops. This study expected...
that a comparison of the residents’ retrospective with their postworkshop self-assessments would yield the same indication of workshop effectiveness as a comparison of the raters’ pre- and postworkshop scores. This was, in fact, what was found. The comparison of the aggregated scores given by the raters and the residents showed that the residents demonstrated more of the target feedback skills after the 2 half-day workshops than they did before the workshop and that this difference was not likely the result of chance. The raw scores and the effect sizes for self-assessments and trained raters are, however, not exactly the same.

The self-assessments yielded a moderately greater effect size than the scores obtained from the trained raters, though they were both relatively close (.95 and .73 out of a usual range of +/−1). If this workshop had been evaluated entirely on the basis of the self-assessments, the organizers would have concluded that their intervention had been very highly successful and had produced a large effect of crucial practical importance. The conclusion would have been somewhat different if based on the scores of trained raters: Organizers might have thought that this workshop had been very successful with high practical importance. Although these two conclusions are similar, caution is advised when making judgments and decisions on self-assessment data only.

The response-shift bias was not detected in this study: There was only a negligible difference between the retrospective self-assessment and preworkshop self-assessment. For this study, a comparison of the pre- and postworkshop self-assessments would have created a reasonable indicator of workshop success. This will not always be the case (Pratt et al., 2000). This workshop likely did not change what the residents knew about good feedback skills but did contribute to their skill development. On the other hand, as was noted in the Data Analysis section, the reliability of the preworkshop self-assessment data was only moderate (Cronbach’s alpha of .59), whereas that of the retrospective self-assessment data was strong (.84). It is plausible that after learning from the workshop, the residents became more reliable and consistent self-assessors—though still somewhat inaccurate! It would be best to routinely collect retrospective, along with postworkshop, self-assessment data as these data will likely be at least as good as (and sometimes better than) the traditional pre- and postworkshop self-assessment data.

There was a lack of correlation between the scores from the trained observers and the self-assessments of individual residents ($r = −.10$). This suggests that many of the residents in this study were poor judges of their own feedback skills, either under- or overassessing their own skill level. This lack of self-awareness is quite common, as noted earlier in this article. This finding might deter workshop evaluators from relying on self-assessment data. Although the individual responses appear to be inaccurate compared with the trained raters, the usefulness of self-assessment data for workshop evaluation rests with the finding that the aggregated grouped data do give an indication of how the group improved as a result of the workshop. Pratt et al. (2000) found that people in the lower portion of the group tended to overestimate their scores and those at the higher end tended to underestimate theirs. This is what happened in this study. Table 2 shows that among those in the bottom third of the group, most overassessed themselves and among the top third, most underassessed themselves. This demonstrates why there was no statistically significant difference between the residents’ self-assessments and the raters’ scores: Overestimates were balanced by underestimates. Although grouped retrospective and postintervention self-assessments have potential to be used in workshop evaluation, even recent recommendations have been cautious (Pratt et al., 2000). Possible limitations include recall and demand bias (Bamberger et al., 2003; Sprangers & Hoogstraten, 1991) as well as other threats to validity (Gall et al., 2003; Rossi, Freeman, & Lipsey, 1999).
Further research would be helpful to enhance the construct validity of aggregated grouped self-assessment data to accurately reflect workshop effectiveness. In this study and in others (Hyman, 1993; Pratt et al., 2000; Skeff et al., 1992), there was congruence between the comparisons of self-assessment data and more objective and rigorously collected data for the purpose of workshop evaluation in situations in which the workshops were judged to be successful. We do not know whether the self-assessment data will correctly identify target skills that were not attained. It is possible that participants might think that they had acquired the knowledge or skills when, in fact, they had not. A possible future test of the validity of self-assessment data for workshop evaluation would be to offer a workshop proven to improve some skills but not others. Evaluators might be more convinced of the efficacy of grouped self-assessment data if it were possible to demonstrate that some skills had been improved and others had not and that trained and blinded raters validated these findings. A study that demonstrated that aggregated grouped self-assessments distinguished between effective and ineffective aspects of a workshop would be very convincing.

Because this is the first time that effect sizes have been compared to establish the construct validity of aggregated self-assessment workshop evaluations, it is not known how closely the effect sizes generated from different data sources might be expected to match. It needs to be determined whether effect sizes derived from self-assessments will always be higher than those from the trained raters and what factors (the nature of the target skills, the type of workshop, the characteristics of the participants, etc.) might contribute to a greater or lesser difference between effect sizes. This would likely increase the sophistication of the use of self-assessments and the confidence that evaluators would have in using self-assessment data.

The congruence of both the effect-size calculations and tests of statistical significance provides reasonably compelling evidence that using aggregated group self-assessment data is a valid method to evaluate this workshop for medical residents and could be relied on to provide reasonably accurate information about workshop effectiveness in other situations.

References

<table>
<thead>
<tr>
<th>Category</th>
<th>Underestimated</th>
<th>Same</th>
<th>Overestimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top third</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Middle third</td>
<td>6</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Bottom third</td>
<td>3</td>
<td>0</td>
<td>11</td>
</tr>
</tbody>
</table>

Note: Thirds were created based on the scores from the trained raters.


