International Journal for the Scholarship of Teaching and Learning

Volume 6 | Number 2 Article 6

7-2012

Efficiency in Assessment: Can Trained Student Interns Rate Essays as Well as Faculty Members?

Tracy L. Cole
Arkansas Tech University, tcole7@atu.edu

Loretta Cochran
Arkansas Tech University, lcochran@atu.edu

Kim Troboy
Arkansas Tech University, ktroboy@atu.edu

Efficiency in Assessment: Can Trained Student Interns Rate Essays as Well as Faculty Members?

Abstract

What are the most efficient and effective methods in measuring outcomes for assurance of learning in higher education? This study examines the merits of outsourcing part of the assessment workload by comparing ratings completed by trained student interns to ratings completed by faculty. Faculty evaluation of students' written work samples provides the most detailed, actionable data useful for improving the curriculum. While this approach may be efficacious, it is also labor-intensive. Both the faculty and student interns were trained to use a scoring rubric developed for this assessment to rate undergraduate student essay responses to an ethical reasoning scenario. The convergent validity, discriminant validity, and source bias showed no significant difference between the values for the student raters versus those for the faculty raters. These findings support the hypothesis that trained student interns can do as well as faculty at evaluating undergraduate student work samples.

Keywords

Assurance of learning, Assessment, Ratings, Evaluation, Outsourcing

Efficiency in Assessment: Can Trained Student Interns Rate Essays as Well as Faculty Members?

Tracy L. Cole
tcole7@atu.edu

Loretta F. Cochran
lcochran@atu.edu

L. Kim Troboy
ktroboy@atu.edu

David W. Roach Arkansas
Tech University
Russellville, Arkansas, USA
droach@atu.edu

Abstract

What are the most efficient and effective methods in measuring outcomes for assurance of learning in higher education? This study examines the merits of outsourcing part of the assessment workload by comparing ratings completed by trained student interns to ratings completed by faculty. Faculty evaluation of students' written work samples provides the most detailed, actionable data useful for improving the curriculum. While this approach may be efficacious, it is also labor-intensive. Both the faculty and student interns were trained to use a scoring rubric developed for this assessment to rate undergraduate student essay responses to an ethical reasoning scenario. The convergent validity, discriminant validity, and source bias showed no significant difference between the values for the student raters versus those for the faculty raters. These findings support the hypothesis that trained student interns can do as well as faculty at evaluating undergraduate student work samples.

Keywords: assurance of learning, assessment, ratings, evaluation, outsourcing

Introduction

As the expectation of assurance of learning continues to be a major theme for accredited colleges and universities, faculty are searching for effective and efficient methods of data collection and evaluation. The most efficient methods usually involve student responses to objective questions (multiple-choice, true/false, etc.) on exams which are embedded in courses. While this approach can be effective for evaluating content knowledge, using open-ended questions that are effective for evaluating complex cognitive processes can be much more difficult. Yet faculty need to evaluate these cognitive processes, such as those involved in problem-solving, critical thinking, moral reasoning, and quantitative reasoning.

Essay responses can capture students' cognitive processing skills, but evaluating such responses is generally time-consuming, and faculty evaluations of student responses can be inconsistent. Use of rubrics and rating scales to measure process skills on several

https://doi.org/10.20429/ijsotl.2012.060206

dimensions can make the evaluations more consistent and accurate (Knight, Allen, & Tracy, 2010). Data from such an analysis can more easily pinpoint common student difficulties at particular stages of the process and therefore suggest areas in which curriculum revision might be needed. However, reading and rating essays carefully takes time.

This study looks at whether the evaluation of these process skills can be outsourced to trained graduate or upper-division students, who are often used in grading undergraduate student work. If these student workers are majoring in fields such as education or college service personnel, this rating work is representative of some of the functions that they will perform in their professional future. While students majoring in other fields may not need to grade essays in their future careers, managers and professionals in many fields need to evaluate employees, processes, products, and written documents based on established sets of criteria. This approach would therefore benefit the student interns as well as save faculty time for activities which usually cannot be outsourced, such as research and curriculum development.

In this experiment, undergraduate students wrote essay responses to scenarios in which an ethical reasoning problem was presented. Both faculty members and student interns were trained to rate these essays on several dimensions using a scale and rubric. The student interns then rated the entire set of essays, while the faculty members rated a randomly selected subset of those essays. The ratings were then compared for convergent validity, discriminant validity, and source bias to determine if the faculty and student intern ratings were similar and how well the ratings identified differences in students' cognitive processing skills on several dimensions.

Background

Rating Accuracy

Much of the literature on accuracy focuses on calculating the correlation between ratings and some known true score, such as an objective outcome measure or rating by an expert (Cronbach, 1955). A primary concern is that the ratings of student essay responses accurately reflect the students' cognitive processing characteristics. If these ratings are not accurate, the basic methodology will not shed light on student learning and accomplishment, and attempting to determine whether student interns could complete such ratings as well as faculty would therefore be pointless.

Lee (1985) suggests that in situations in which the process being measured is understood but valid and reliable performance measures are not available, the use of behavior-based rating scales will provide more accurate ratings. Where ratings are more subjective and have less specific objective measures of desirable outcomes with which to compare, Borman (1979) found that rater training and a behavior-based rating scales format positively impacted rater accuracy by reducing some rating errors. His example of a task that has specific objective measures was that of evaluating recruiters; the number of successful recruits provides a specific objective measure with which to compare more subjective job evaluation ratings. Evaluating managers, on the other hand, is a situation in which the job outcomes are more subjective and cannot as easily be used to measure the accuracy of job evaluation ratings.

Benson, Buckley, and Hall (1988) examined the impact of rating scale format on rater accuracy. They found that ratings of video-taped interviewer performances were more accurate when the rating scales were behaviorally anchored as compared to a mixed standard scale format. They go on to suggest that the way anchors are defined can affect the accuracy of ratings, stating that "Barnes, Farrell, and Weiss (1984) found that using extreme anchors for mixed standard scales resulted in a reduction of logical inconsistency errors" (p. 418).

Rater Reliability

Researchers generally agree that any method of rating or measuring a product must provide consistency of measurement (Knight et al., 2010). In academic assessment, the reliability of ratings using rubrics involves two aspects: inter-rater reliability or "reproducibility," which refers to the consistency of scores assigned by different raters, and intra-rater reliability or "repeatability," which refers to consistency of scores assigned by the same rater to the same subject at different points in time (Knight et al., 2010; Moskal & Leydens, 2000). As stated by Moskal and Leydens, we should "expect to attain the same score regardless of when the student completed the assessment, when the response was scored, and who scored the response" (Reliability section, para. 1).

While essays or open-ended responses to questions provide a broad measure of students' knowledge or reasoning processes, these instruments present obvious problems for obtaining reliable ratings based on subjective criteria. Scoring rubrics address this problem by formalizing or defining the criteria on which an essay is evaluated. Thus, Moskal and Leydens (2000) contend, a well-designed scoring rubric can increase inter-rater reliability in assessing students' written work. Even so, Knight et al. (2000) noted that some studies have shown relatively poor inter-rater agreement.

This study focuses on inter-rater reliability rather than intra-rater reliability, as each rater evaluated each essay only once. While it is important to obtain consistent results by each individual rater, the purpose of this investigation was to determine whether trained graduate and upper-division students could effectively serve as a proxy for faculty members in rating essays or other written responses. In other words, can we achieve reliable results between these two types of raters?

Training Raters

Some researchers have found that training raters in certain situations does not improve rater accuracy. Upon reviewing the literature, Lee (1985) concluded that efforts to train raters in the use of performance appraisal instruments and procedures have not resulted in increased rating accuracy. Moreover, Benson et al. (1988), citing a study by Bernardin and Pence (1980), stated that training raters to avoid certain psychometric rating errors resulted in situation-specific responses that actually reduced the accuracy of performance evaluations.

In general, however, the literature provides ample support for the proposition that proper training can improve rater accuracy. Tziner (1984) noted that "it is frequently demonstrated in the literature that training affects rater observational and evaluative skills, thus improving rater accuracy" (p. 104). For example, Thornton and Zorich (1980) found that behavioral training alone increased rater accuracy, and the addition of training to avoid systematic errors of observation, such as prejudice and stereotyping, categorization error, and contamination from prior information, further enhanced the accuracy of behavioral observations.

https://doi.org/10.20429/ijsotl.2012.060206

Specifically, frame-of-reference training, which was employed in the present study, has been found to produce more accurate ratings and improve inter-rater reliability (Bernardin & Buckley, 1981; Lee, 1985; Pulakos, 1984). Citing Bernardin and Beatty (1984), Lee explained that “frame-of-reference training is designed to reduce arbitrary performance standards by having raters discuss their own standards in comparison with the normative standards” (p. 328). Through group discussion and problem-solving, the raters can agree on performance criteria before the evaluation process begins, resulting in higher inter-rater reliability.

Upon reviewing the literature, Bernardin and Buckley (1981) concluded that most rater training programs had previously focused on changing rater response distributions and that this approach was not effective for improving rater accuracy. They argued that rater training programs should instead emphasize frame-of-reference training. Upon considering the cognitive processes involved in the rating task, including attention, categorization, recall, and information integration, Pulakos (1984) contended that the most important component is categorization, because it links the other processes together. Thus, Pulakos suggested, training that focuses on creating or imposing a system of categorization, such as frame-of-reference training, may be more useful for improving accuracy than training designed to reduce psychometric errors.

Methodology

The authors of this study coordinated an ethics assessment activity during the 2007-2008 academic year. Both upper- and lower-division undergraduate business students participated in a graded essay assignment, either as part of their course requirements or for extra credit. Collecting responses from students at the beginning and at the end of their academic programs allowed some consideration of whether the college experience made any difference in the students’ ethical decision-making process. Students were assured of individual confidentiality of their responses.

The scenario chosen for this assessment was *Gilbane Gold* (National Institute for Engineering Ethics, 1989). This scenario is available in both video and script format. This hypothetical case describes a dilemma faced by a character named David Jackson, an engineer at the Z-Corp manufacturing plant. He must deal with the measurement of the level of toxic waste that Z-Corp discharges into a city’s sewer system. The city sells sludge from that sewer system to farmers who use it for fertilizer, and the city then uses the income from these sales to lower the tax burden on citizens. The possible outcomes include consequences such as threats to Jackson's license and job, company profitability, plant viability, public health, city revenues, and taxpayer burden. The scenario was chosen because it presents a complex ethical dilemma in a business context.

Phase One of Data Collection - Essay Responses

One hundred seventy-six undergraduate students participated in the assessment. The students first filled out a demographic questionnaire and then either watched a video version or read a transcript of the *Gilbane Gold* scenario. Data analysis showed no significant differences with respect to which format the students were exposed. All participants were asked to answer the following short-answer questions in essay format to elicit multiple aspects of ethical awareness and decision-making:

1. What are the goals and objectives David Jackson should consider?

2. What alternatives should David Jackson consider?
3. What should David Jackson do and why?
4. Who will be affected by this decision?

The authors developed these questions based on a pilot study conducted in the fall of 2007 (Cochran, Roach, Troboy, & Cole, 2010).

The students completed the exercise in one hour or less in a proctored classroom setting. They were not were not alerted to address possible ethical issues because one dimension of moral reasoning being measured was the ability to identify the existence of a moral dilemma without being prompted to do so.

Phase Two of Data Collection - Student Intern and Faculty Ratings of Essay Responses

The essay responses were coded and presented in such a fashion that the raters could not determine which student wrote an essay or whether the student was enrolled in an upper- or lower-division course. Each student response sheet was labeled with a randomly generated identification code on the name/demographic sheet and the short-answer question response sheet. These two sheets were separated and the short-answer question response sheets were sorted into numeric order. Copies of the re-ordered short-answer question response sheets were distributed to the raters. This approach limited the ability of the raters to guess the identity of the respondents.

The authors had previously developed a rating rubric with specific, concrete anchors for categorizing and scoring the essays (Cochran et al., 2010; Cole, Cochran, Troboy, Roach, & Wu, 2008). The rubric consisted of four dimensions of ethical reasoning rated on an 8-point scale. The four dimensions of ethical reasoning included:

1. Identifies dilemma.
2. Considers stakeholders.
3. Identifies and analyzes alternatives.
4. Identifies ethical outcomes and consequences.

These dimensions were not designed to correspond directly to the short-answer questions used on the assessment instrument. Instead, these four issues represent an organized framework for ethical decision-making and are readily recognized in student responses. They are also teachable approaches to solving problems in professional situations.

The eight points on the scale were grouped by twos into four categories: unacceptable, marginal, acceptable, and exemplary. The authors initially tried using a 4-point scale, so that each category corresponded to one point on the scale. However, this approach proved insufficient to differentiate the respondents' levels of proficiency in ethical reasoning. The scale was therefore expanded, resulting in the 8-point scale and anchored rubric shown in Appendix A.

Three faculty members participated as expert raters. The four interns participating in the study as raters included three graduate students from a Master's program in college student personnel and one undergraduate who was a senior in a business administration program. The sample size for this study was therefore 176 papers rated by four interns and 49 papers rated by three faculty members. Ideally, a study such as this one might include a larger number of student interns. A previous study by Roach, Lucas, Cole, Braunsberger, and

Bequette (2000), however, found that a larger number of student raters (29) could achieve results similar to five faculty members when rating a very small number of writing samples (3). For this study, the authors chose to use a smaller number of raters evaluating a much larger sample of papers, as this model reflects the circumstances under which assessment activities would be conducted in most academic settings.

To make sure that all raters were in agreement on meaning and use of the rating rubric, the authors conducted an initial frame-of-reference training session for the raters using three sample essays (Pulakos, 1984; Pulakos, 1986). The training session was intended to standardize and improve the accuracy and validity of the rating process.

Before the training, the raters were given the script of *Gilbane Gold* to read. One of the authors, a senior faculty member in management, then led an overview and discussion of the case in relation to the rubric and rating criteria. The evaluation criteria for each dimension of ethical reasoning were discussed and expanded. Then each rater independently scored three sample essays. After each sample essay was scored, the raters' individual scores were displayed on a whiteboard, and the raters discussed why they assigned particular scores and what criteria they believed exemplified various performance levels. The sample essays and corresponding discussions were thus completed in three rounds, with each one bringing the raters more closely into agreement on appropriate scores. Altogether, the training session took about an hour and a half to complete.

This frame-of-reference training served two purposes. First, it compensated for any lack of practical or academic experience among the student interns. Second, it provided a means of calibrating the raters, enabling ratings to be fixed with respect to a standard of performance (Sulsky & Balzer, 1988).

Data Analysis

In order to get a "true" measure of rater accuracy, three faculty members independently rated 49 randomly selected papers from the 176 total papers collected for the study. The student interns rated all 176 papers. Both the faculty and student ratings were validated by computing indices of convergent validity, discriminant validity, and source bias, which is the approach proposed by Kavanagh, MacKinney, and Wolins (1971). The results are shown in Table 1 below.

Convergent validity refers to the level of agreement between multiple measures of a particular trait in a multitrait-multimethod (MTMM) matrix (Campbell & Fiske, 1959). Kavanagh et al. (1971) reframed convergent validity in terms of multitrait-multirater (MTMR) matrix and developed a single, interpretable index to summarize convergent validity, where convergent validity gages the level of agreement between multiple raters of a particular dimension of rater performance.

Discriminant validity refers to the degree to which traits that should differ (theoretically) are, in fact, not related when those traits are measured (Campbell & Fiske, 1959). Kavanagh et al. (1971) reframed discriminant validity in terms of a MTMR matrix and developed a single, interpretable index to summarize discriminant validity, where discriminant validity gages the degree to which the various rated dimensions of performance are, in fact, unique or separate dimensions.

In addition to providing measures for convergent and discriminant validity, Kavanagh et al. (1971) provided an index that indicates the source bias of performance ratings. Source bias gages the degree to which a rater's evaluation is affected by his or her individual, subjective view of performance.

The multirater feature of Kavanagh et al.'s (1971) MTMR matrix refers to different "types" of raters: supervisor, peer, and self. Borman (1978) reframed the MR portion of the matrix as multiple "expert" raters. In this study, the authors compare characteristics (convergent validity, discriminant validity, and source bias) of ratings of two "types" of experts: faculty who have completed frame-of-reference training and upper division or graduate students who have completed frame-of-reference training.

Table 1

Comparison of Validity Measures Between Faculty and Student Interns

	Faculty[1]	Trained Student Interns[2]	p-value for difference between correlations[3]
Convergent Validity	0.66	0.63	.7566
Discriminant Validity	0.24	0.20	.8026
Source Bias	0.35	0.41	.6745

[1] Three faculty members each rating 49 papers.

[2] Four raters (three graduate students and one senior) each rating 176 papers.

[3] Based on Fisher r to z transformation and a two-tailed test.

The convergent validity index for faculty and students was .66 and .63 respectively. The discriminant validity index for faculty and students was .24 and .20 respectively. These results fall within the range of previous performance appraisal studies (e.g., Borman, 1978; Roach, 1991; Roach & Gupta, 1990; Roach & Gupta, 1992; Roach et al., 2000).

The source bias was 0.35 and .41 for each group. The source bias is higher than (not as good as) previous studies where performance was scripted (Borman, 1978; Roach et al., 2000) but lower (better) than a study, like that employed in this study, where performance was not scripted (Roach & Gupta, 1992).

The results described above suggest that both the faculty members and student interns provided ratings with sufficient validity to warrant their use as estimates of true performance. The most important findings, however, are the p-values for the differences between the student intern and faculty ratings. The p-values (all > .2) support the idea that no significant differences exist between the results for faculty versus trained student raters, which shows strong inter-rater reliability for the two groups.

Conclusions

The most significant finding of this study is that trained student raters can achieve essentially the same results as faculty members when assessing written work product of undergraduate students. Moreover, both the faculty and student raters achieved a reasonable level of convergent validity in their ratings. Though discriminant validity was lower than that reported in previous studies, this result may reflect true correlation between the dimensions of ratee performance. The use of a grading rubric along with frame-of-reference training to ensure that raters understand and agree on the grading criteria has previously been shown to increase rater accuracy. This study suggests that this method can be employed with student raters, as well as faculty or expert raters, to achieve valid results.

Colleges and universities increasingly find themselves under pressure to conduct more thorough assessments of student learning, while at the same time struggling with limited institutional resources and greater demands on faculty members' time and attention. While the approach used in this study requires funds to pay student raters, student labor is far less expensive than faculty labor. Thus, a well-designed assessment using student raters may both enable institutions to conduct high quality assessment initiatives and allow faculty members to better use their time and expertise for research, curriculum development, and other academic functions.

References

Benson, P. G., Buckley, M. R., & Hall, S. (1988). The impact of rating scale format on rater accuracy: An evaluation of the mixed standard scale. *Journal of Management*, *14*(3), 415-423.

Bernardin, H. J., & Beatty, R. W. *Performance appraisal: Assessing human behavior at work*. Boston, MA: Kent.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *The Academy of Management Review*, *6*(2), 205-212.

Bernardin, H. J., & Pence, E. C. (1980). Rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology*, *65*(1), 60-66.

Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology*, *63*(2), 135-144.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, *64*(4), 412-421.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105.

Cochran, L., Roach, D., Troboy, K., & Cole, T. (2010). Developing an essay approach to business ethics assessment. *The Journal of American Academy of Business, Cambridge*, *15*(2), 37-42.

Cole, T., Cochran, L., Troboy, K., Roach, D., & Wu, C. (2008). Refining a measure of ethical reasoning and decision-making. *Proceedings of the Academic Business World International Conference*, Nashville, TN, 470-480. Retrieved from <http://abwic.org/>

Cronbach, L. J. (1955). Processes affecting scores on 'understanding of others' and 'assuming similarity'. *Psychological Bulletin*, *52*(3), 177-193.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin*, *75*(1), 34-49.

Knight, J. E., Allen, S., & Tracy, D. L. (2010). Using six sigma methods to evaluate the reliability of a teaching assessment rubric. *The Business Review, Cambridge*, *15*(1), 1-6.

Lee, C. (1985). Increasing performance appraisal effectiveness: Matching task types, appraisal process, and rater training. *Academy of Management Review*, *10*(2), 322-331.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10). Retrieved from <http://pareonline.net/>

National Institute for Engineering Ethics. (1989). *Gilbane Gold: An Ethics Story Focusing on Responsibilities of Engineers*. Retrieved from <http://www.murdough.ttu.edu/pdg.cfm?pt=NIEE&doc=ProductsServices-GilbaneGold.htm>

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology*, *69*(4), 581-588.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes*, *38*, 79-91.

Roach, D. (1991). *An investigation of rating accuracy in a realistic setting*. (Unpublished doctoral dissertation). University of Arkansas, Fayetteville, AR.

Roach, D., & Gupta, N. (1990). Contextual effects on rating leniency: A realistic simulation. *Proceedings of the 41st Annual Meeting of the Academy of Management*, San Francisco, CA.

Roach, D., & Gupta, N. (1992). A realistic simulation for assessing the relationships among components of rating accuracy. *Journal of Applied Psychology*, *77*(2), 196-200.

Roach, D., Lucas, L., Cole, G., Braunsberger, K., & Bequette, J. (2000). Using undergraduate students to assess business curriculum outcomes. *Proceedings of the Society for Advancement of Management 2000 International Management Conference: Managing in a World of Change*, St. Augustine, FL, 446-451.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and Measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*(3), 497-506.

Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology*, *65*(3), 351-354.

Tziner, A. (1984). A fairer examination of rating scales when used for performance appraisal in a real organization setting. *Journal of Occupational Behaviour*, *5*, 103-112.

Appendix A
Ethical Reasoning Rating Rubric

TRAIT	Unacceptable	Marginal	Acceptable	Exemplary
Identifies Dilemma	No recognition of ethical dimensions of dilemma 1 2	Limited recognition of ethical dimensions of dilemma 3 4	Accurately identifies ethical dimensions of dilemma 5 6	Accurately identifies & examines (with contemplation) ethical dimensions 7 8
Considers Stakeholders	No consideration of stakeholders 1 2	Limited consideration of stakeholders 3 4	Accurately identifies most or all stakeholders 5 6	Thoroughly contemplates viewpoints of most or all stakeholders 7 8
Identifies & Analyzes Alternatives	Lists one option with little or no evaluation 1 2	Lists two alternatives with little or no evaluation 3 4	Identifies and lists some pros & cons for two or more alternatives 5 6	Accurately identifies & examines ethical dimensions (from >2 stakeholders) 7 8
Identifies Ethical Outcomes & Consequences	Proposes an unethical course of action 1 2	Fails to identify ethical outcome &/or consequence 3 4	Limited identification of an ethical outcome or consequence 5 6	Clearly identifies & examines one or more ethical outcomes &/or consequences 7 8