RELIABILITY OF SINGLE-ITEM RATINGS OF QUALITY IN HIGHER EDUCATION: A REPLICATION¹

PAUL GINNS AND SIMON BARRIE

University of Sydney

Summary.—Single-item ratings of the quality of instructors or subjects are widely used by higher education institutions, yet such ratings are commonly assumed to have inadequate psychometric properties. Recent research has demonstrated that reliability of such ratings can indeed be estimated, using either the correction for attenuation formula or factor analytic methods. This study replicates prior research on the reliability of single-item ratings of quality of instruction, using a different, more student-focused approach to teaching and learning evaluation than used by previous researchers. Class average data from 1,097 classes, representing responses from 59,815 students, were analysed. At the "class" level of analysis, both methods of estimation suggested the single item of quality had high reliability: .96 using the correction for attenuation formula, and .94 using the factor analytic method. An alternative method of calculating reliability, which takes into account the hierarchical nature of the data, likewise suggested high estimated reliability (.92) of the single-item rating. These results indicate the suitability of the overall class rating for quality improvement in higher education, with a large sample.

Single-item indicators of a theoretical construct have mixed support in the behavioural sciences. While such indicators are often used in applied research, such as surveys of customers’ satisfaction and evaluations of employees’ performance, researchers aiming to publish their research tend to measure such constructs using multiple indicators. Wanous and Hudy (2001) noted many researchers believe that the reliability of a single item cannot be estimated, and that even if it could, that it would be unacceptably low. The presumption is that multiple indicators are necessary to measure variability of latent constructs.

However, when the construct of interest is relatively narrow or is unambiguous to respondents, a single-item measure may be appropriate (Wanous, Reichers, & Hudy, 1997). Novel applications of psychometric theory and factor analytic methods have made possible such estimation focusing on overall ratings of various types of satisfaction. Wanous and Reichers (1996) showed how the widely used correction for attenuation formula could be used to estimate single-item reliability. Wanous, et al. (1997) applied this method to estimating the reliability of single-item measures of Overall Job Satisfaction.

¹Address correspondence to Dr. Paul Ginns, Institute for Teaching and Learning, 3rd Floor Carslaw Building F07, University of Sydney, NSW 2006, Australia or e-mail (p.ginns@fcl.usyd.edu.au).
Using meta-analytic methods to analyse 28 correlations based on 7,682 persons’ ratings, they concluded a minimum single-item reliability close to 70 was feasible. However, Loo and Kells (1998) did not replicate this finding when examining survey responses on Job Satisfaction from a sample of teaching and administrative staff (N=562) at a community college, finding a single-item reliability for Overall Job Satisfaction of only .43.

Wanous and Hudy (2001) extended this research by estimating the reliability of single-item measures of College Teaching Effectiveness, using an additional method, factor analysis, to estimate single-item reliability. Their results indicated single-item reliability estimates were higher using factor analysis (.88) than the correction for attenuation formula (.64), and were higher using class-level data (.82) than individual-level data (.70). These methods for estimating such reliabilities are reviewed below.

**Estimating Single-item Reliability with the Correction for Attenuation Formula**

The correction for attenuation formula is a long-standing component of classical psychometric theory. The formula can be written as follows (cf. Nunnally, 1978, p. 220):

$$r_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} \cdot r_{yy}}}$$

where $r_{xy}$ is the correlation between two variables $x(1)$ and $y(2)$, $r_{xx}$ is the reliability of variable $x$, $r_{yy}$ is the reliability of variable $y$, and $r_{xy}$ is the estimated true correlation between $x$ and $y$ if the two variables were measured with perfect reliability. Typical applications of the formula arise when the variables are drawn from separate domains, such as “Self-efficacy” ($x$) and “Training Performance” ($y$). However, when both variables are drawn from the same domain, Nunnally (1978, p. 220) noted “[this equation] also applies to samples of items drawn from the same domain [emphasis in original]... the correlation between two such tests would be expected to equal the product of the terms in the denominator and consequently $r_{xy}$ would equal 1.00... If $r_{xy}$ were 1.00, $r_{xx}$ would be limited only by the reliability of the two tests: $$r_{xy} = \sqrt{r_{xx} \cdot r_{yy}}$$

The use of this method to estimate single-item reliability can be demonstrated using Wanous and Hudy’s class-level results. (Class-level results were the main focus of their study, as the university from which the teaching evaluation results were derived only used class means for the purpose of quality assurance.) They estimated the reliability of a nine-item scale of Teaching Effectiveness to be .88, and the correlation of this score with the overall rating of Teaching Effectiveness to be .79. Solving the above equation for the missing value of $r_{xx}$ gives a minimum estimated single-item reliability of .71.

**Estimating Reliability with Factor Analysis**

Common factor analysis provides an alternative method for calculating single-item reliability. Harman (1967, pp. 16-19) noted that the variance of an item can be represented as:

$$\text{Total Variance} = \text{Communality} + \text{Specificity} = \text{Unreliability}$$

The reliable variance of an item is therefore the sum of its communality and its specificity, making the communality less than or equal to the reliability of that item. If there is no specific variance, the communality equals the reliability; that is, the communality of an item may be used as a conservative estimate. Arvey, Landon, Nutting, and Maxwell (1992) used this estimation method as part of a larger research project validating tests of physical ability for police officers, deriving single-item reliabilities between .61 and .73 for a range of physical ability tests. Using this method on ratings of Teaching Effectiveness, Wanous and Hudy (2001) estimated the overall rating single-item reliability as .94 for group level data, and .81 for individual level data.

**The Present Study**

The present study replicates Wanous and Hudy’s analyses (2001) by investigating the single-item reliability of university students’ ratings of overall satisfaction with an individual class. It extends their analyses by examining students’ ratings collected at a different institution, in a different country, and using a scale based on a different model of teaching and learning evaluation (student-focused) to that forming the basis of Wanous and Hudy’s scale (teacher-focused; for a discussion of this distinction, see Prosser & Trigwell, 1999). In addition, we compared reliability estimates computed using the above method with those calculated using another method, which takes into account the hierarchical nature of the data (students’ ratings nested within classes). This method, described by Snijders and Bosker (1999, p. 26), expresses the reliability, $\lambda$, of an aggregated variable as

$$\lambda = \frac{n_{ij} \cdot p_{ij}}{n_{i} \cdot p_{ij} + (n_{i} - 1) \cdot p_{ij}}$$

where $n_{ij}$ equals the average cluster (class) size, and $p_{ij}$ equals the intraclass correlation (the average correlation between overall ratings made by students in the same class). This method of calculating reliability is an extension of classical test theory, which defines reliability as the variance of true scores divided by the variance of observed scores (Snijders & Bosker, 1999). Typical applications of the notion of reliability refer to the average consistency of responses across a set of related items, e.g., in a scale. The above method
estimates the average consistency of ratings in the case where several single-item measurements are made on a number of macro-level units, e.g., a class. Snijders and Bosker (1999, p. 23) also provide a formula for estimating the standard error of the intraclass correlation:

\[ SE(p_c) = \frac{(1-p_c)1-(n-1)p_c}{\sqrt{2(n)(n-2)}} \]

where \( n \) equals the average cluster (class) size, \( p_c \) equals the intraclass correlation, and \( N \) equals the number of clusters. An approximate 95% confidence interval for the intraclass correlation may be calculated as the intraclass correlation plus or minus 1.96 times the estimated standard error. In turn, the upper and lower values of the confidence interval may be substituted into the reliability formula given above to derive an approximate 95% confidence interval for the reliability estimate (T. Snijders, personal communication, August 25, 2004). Methods for calculating reliability coefficients’ confidence intervals for the correction for attenuation and factor analytic methods could not be located. For further discussion of methods of estimation of confidence intervals around reliability coefficients, see Pan and Thompson (2001).

**Method**

**Sample**

The data set consisted of students’ ratings of undergraduate classes collected over a two-year period (2001–2002) at the University of Sydney, Australia. Ratings were collected from 1,097 individual classes taught on the main campus and at two branch campuses. Of a total of 103,190 students enrolled in these classes, 59,813 responded on the Unit of Study Evaluation (USE), giving a total response rate of 57.97%. As in Wanous and Hudy’s analysis (2001), classes with fewer than five students’ ratings were excluded from analysis, to avoid the possibility of mean ratings being unduly skewed by a single student’s rating. The group level of analysis is the primary focus of this study, as it was for Wanous and Hudy, since the University uses this level of analysis, i.e., mean ratings, for purposes of quality assurance. While reliability is commonly assessed using data collected at the level of the individual, such methods ignore the portion of error due to lack of agreement between individual respondents (Gilmore, Kane, & Naccarato, 1978). For this reason, in his comprehensive review of research on students’ evaluations of teaching, Marsh (1987) argued that the reliability of the class-average response is the appropriate method for assessing reliability.

**Measures**

The Unit of Study Evaluation (USE) is intended as a source of data to support improvement in the quality of teaching and learning, with seven core items regarding the student’s experience and one overall rating of satisfaction with the quality of the class (referred to within the University as a “unit of study”). The USE items reflect a student-focused learning perspective. Research conducted within this educational paradigm has demonstrated that students’ approaches to study are contingent upon both their prior experiences and perceptions of current learning contexts (e.g., Prosser & Trigwell, 1999; Ramsden, 2003). Researchers investigating the effects of such perceptions in a wide range of educational domains have found students who adopt a surface approach tend to have poorer learning outcomes than those adopting a deep approach (see Prosser & Trigwell, 1999, Chap. 6, for further discussion). Thus, a major purpose of the USE is to assess the extent to which students perceive the class as engendering deep approaches to learning. In contrast to other systems of Teaching Evaluation, the focus of the USE is not the qualities of the instructor, i.e., a teacher focus, but the student’s experience of the class, as often there is more than one person with teaching responsibilities for a given class. Other measures for providing feedback to teachers are available for more personalized feedback.

Each item in the USE uses a 5-point Likert scale with five anchors, 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree, and space is given under each item for students to explain the reasons for their rating, provide suggestions for improvement, or both. The seven items are (1) the learning outcomes and expected standards of this unit of study were clear to me; (2) the teaching in this unit of study helped me to learn effectively; (3) this unit of study helped me develop valuable general attributes (e.g., written and oral communication skills, analytic skills, information literacy, teamwork skills, etc.); (4) the workload in this unit of study was too high (reversed); (5) the assessment in this unit of study allowed me to demonstrate what I had understood; (6) I can see the relevance of this unit of study to my degree; and (7) it was clear to me that the staff in this unit of study were responsive to student feedback. The wording of the Overall Quality rating item is “Overall I was satisfied with the quality of this unit of study.” To ensure confidentiality, students are not required to provide any identifying information, e.g., a student number or demographic information.

USE data are collected for nominated classes in the twelfth and thirteenth weeks of a 14-wk. semester. Students complete scannable forms, with a nominated student representative being responsible for collecting the forms, sealing them in an envelope, and returning them to the class coordinator who waits outside the classroom during this process. Forms are returned to the Institute for Teaching and Learning for scanning and report generation, with results being returned to class coordinators and deans of faculties. Data for the present study were extracted from a database of student responses.
The scores for individual USE items (excluding the overall rating) were combined into an average score (referred to by Wanous and Hudy (2001) as a facet score) in the following way. Initially, all cases nested within classes where there were fewer than five cases were excluded. For individual level data, the data were screened for missing values on the overall rating, and all cases in which this rating was missing were excluded. Ratings on the other seven items were combined, using an SPSS algorithm that creates an average score if ratings were given on more than half, i.e., at least 4 out of 7, of the items. Any cases for which an average rating could not be created using this algorithm were excluded. For group level data, mean scores for all cases in a given class were created for each of the eight USE items from the individual-level dataset (from which low-response data had been excluded), and a mean overall score was created by averaging the means of the group-level items.

**Results**

**Correction for Attenuation Method**

As mentioned above, the most conservative approach to estimating the underlying correlation between the two related constructs, i.e., the average scale score and the Overall Quality rating, is to assume a correlation of 1.00. For the class level data, the Cronbach reliability coefficient of the facet scale was .88, and the correlation between the facet scale and the Overall Quality rating was .92. Assuming an underlying correlation of 1.00, this gives an estimated single-item reliability of .96.

**Factor Analytic Method**

The single-item reliability of the Overall Quality rating was next estimated using principal axis factor analysis, using the communality of the Overall Quality rating as a conservative estimate of single-item reliability. One factor was extracted with an eigenvalue greater than one, accounting for 62.85% of the variance. The communality of the Overall Quality rating was .94. [With a more recent data set, a census of all classes of a specific faculty (N = 268 classes), r was .94.]

**Intraclass Correlation Method**

The intraclass correlation for Overall Quality ratings within classes was calculated as .27. Because the average number of respondents per class (48.7) and the median number of respondents (28) differed, we used the geometric mean (29.4) rather than the arithmetic mean for the average class size. Using the formula from Snijders and Bosker (1999), the reliability of the Overall Quality rating at the class level was estimated as .92 (95% confidence interval .907 – .923).

**Discussion**

The results for both the methods of correction for attenuation and factor analysis for estimating single-item reliability compare well with the results obtained by Wanous and Hudy (2001). At the class level, the primary level of analysis for quality assurance at both universities, the results suggest viable estimates of single-item reliability are possible using either method. At this level, the results of the two studies were the same using the factor analysis method (.94). Using the attenuation formula, Wanous and Hudy estimated a value at the group level of .78, while the present study obtained a value of .96. This higher estimate can be attributed to the higher correlation observed in the present study between the facet scale and the overall rating (.92), compared to Wanous and Hudy's observed correlation of .79 and Loo and Kells' (1998) observed correlation of .65. The items used at Wanous and Hudy's institution focused on evaluating the teacher, including an overall teachers' rating, while the USE focuses on the student's experience, evaluating satisfaction with the overall quality of the class. The higher correlation in the present study may be an indication of the superiority of student-focused items over teacher-focused items in student evaluations, but additional research investigating this system in other institutional contexts would be needed to substantiate this hypothesis. More generally, the utility of the correction for attenuation formula for estimating single-item reliability is constrained by the correlation between the composite score, e.g., Job Satisfaction, Teaching Quality, and the single-item overall rating, a fact that researchers must take into account when considering the use of this method. The intraclass correlation method, which explicitly takes into account the nested nature of the data, also yielded an estimate of reliability (.92) similar in magnitude to the other methods. The narrowness of the confidence interval around this estimate (.907 – .923) also gives us additional confidence in the use of the Overall Quality ratings for quality improvement.

The results of the present study replicate those of Wanous and Hudy (2001), showing that, at a class level, ratings of such indicators correlate well with composite scores based on other evaluation factors. In turn, this indicates the overall rating is an acceptable proxy for a construct describing the overall class quality. The similarity of the different single-item reliability estimates obtained at the class level gives us confidence in the item, "overall satisfaction with quality" as a meaningful source of information, allowing accurate identification of high quality classes that might be used for benchmarking, and classes requiring reassessment of teaching and learning strategies.

**References**


Accepted October 22, 2004.