Copyright © 2013, New Forums Press, Inc., P.O. Box 876, Stillwater, OK 74076. All Rights Reserved.

Should Global Items on Student Rating Scales Be Used for Summative Decisions?

Ronald A. Berk
The Johns Hopkins University

Introduction

One of the simplest indicators of teaching or course effectiveness is student ratings on one or more global items from the entire rating scale. That approach seems intuitively sound and easy to use. Global items have even been recommended by a few researchers to get a quick-read, at-a-glance summary for summative decisions about faculty. So why is there so much confusion, misunderstanding, and debate on professional listservs, articles, and books on whether global items should be used? Also, how many times can you use a series of three nouns in the same sentence? What about adjectives? This topic has become a rather prickly, thorny, and knotty issue that most administrators must tackle (Berk, in press).

The administrative choice is making a decision, such as contract renewal or a pay raise, based on one global item on the professor's performance or the single total rating based on all of the rating scale items. In other words, should your department chair use the rating from one item or a collection of items? Usually the ratings are gathered from all of an instructor's courses taught over a year (annual review) or three or more years (promotion and tenure review). How much information is needed for summative decisions? What are the technical differences in the reliability and validity of these single rating options?

The purpose of this article is to clarify the rating options available to administrators for making summative decisions from student rating scale results. The global item is first defined. Then the research literature related to its use for summative decisions is described. Finally, the critical issues pertaining to technical, professional, and legal standards in using global items for summative decisions are reviewed. Specific recommendations for several alternative types of ratings are proffered which are more defensible for summative decisions about faculty.

Definition of a Global Item

"What in the world is a *global item*?" Here are a couple of examples rated with a "Strongly Disagree" to "Strongly Agree" anchor scale:

- Overall, this instructor was an excellent teacher.
- Overall, this course was excellent.

This type of item provides a general *broad-stroke, summary index of teaching performance or course quality*. It doesn't address specific teaching and course characteristics. Global items typically appear at the end of the student rating scale so students have had time to form an opinion after responding to all of the items. They should not be summed with the ratings of all other items; they are reported separately, independent of the rest of the scale.

Use for Summative Decisions

In the 1990s, administrators, such as department chairs, directors, or associate deans, were encouraged to use the ratings on global items to provide a *simple, quick-and-easy measure of teaching effectiveness for summative decisions* (Abrami & d'Apollonia, 1991; Algozzine et al., 2004; Braskamp & Ory, 1994; Cashin & Downey, 1992; Cashin, Downey, & Sixbury, 1994; Centra, 1993; d'Apollonia & Abrami, 1997; Hativa & Raviv, 1993). More recently, administrators expressed a preference for the use of global items for information on the overall quality of the course and instructor (Beran, Violato, & Kline, 2007; Beran, Violato, Kline, & Frideres, 2005). Those ratings may or may not be used in conjunction with other information to arrive at decisions regarding merit pay, contract renewal for full-time and adjunct faculty, teaching awards, and promotion and tenure recommendations.

What's the Problem?

Despite the simplicity and ease with which an administrator can use a single global item rating for summative decisions about faculty, that use is *inappropriate for personnel decisions about employees*. There are several sets of standards that specify the types of ratings that are required for such decisions. The global item does not satisfy all of those standards. The next section explains why.

Critical Standards

Given the seriousness of these decisions on a professor's condition of employment and career, *these standards should be met by all administrators* who are responsible (Berk, in press). The standards can be sorted into four categories: (1) psychometric, (2) representativeness and fairness, (3) professional, and (4) legal.

Psychometric Standards

Item validity evidence. Cashin and Downey (1992) studied two global items:

- "Overall, I rate this INSTRUCTOR an excellent teacher," and
- "Overall, I rate this COURSE as excellent."

They found that these items accounted for more than 50% of the variance in a composite criterion measure—the *Instructional Development and Effectiveness Assessment* (*IDEA*). When this study was replicated with four other criteria using *IDEA* data (Cashin et al., 1994), the results were the same: *a body of validity evidence that the global items accounted for most of the variance in several criterion measures of teaching effectiveness.* This prompted the researchers to recommend the use of those items for summative decisions.

Although global item-total scale and subscale correlations are rarely reported, there is also evidence that those combinations are highly intercorrelated (Harrison, Douglas, & Burdsall, 2004; Hativa & Raviv, 1993; Otani, Kim, & Cho, 2012). Item intercorrelations between course and instructor global items and items on teaching methods and student progress on course outcomes were consistently moderate to high for the *IDEA Student Ratings of Instruction* form (Benton et al., 2010b). So, with all of this compelling validity evidence, *why not substitute the global rating for the total rating?* Wouldn't that be a reasonable proxy? Not exactly.

Item reliability evidence. The issue is what's not usually computed: the reliability of the global item rating. Rarely are global item reliabilities based on *class means* estimated for student rating scales. Typically, *item reliabilities can be in the .70s–.90s* (Ginns & Barrie, 2004; Wanous & Hudy, 2001) *for unidimensional scales*, depending on the method used to estimate them (test-retest, intraclass correlation, correction for attenuation, or factor analysis). That range of coefficients can be illustrated with the Spearman rho split-half coefficients of .75–.91 (class sizes = 10–34) and the .90s (class sizes = 35 and above) for the global items on the *IDEA Student Ratings of Instruction* form (Hoyt & Lee, 2002).

The problem is that coefficients in the .70s are too low and unstable for single global items to be used for decisions about individual employees. Since those items are usually found on multidimensional scales, the item reliabilities can be even lower. In contrast to these reliabilities, the *reliability coefficients in the mid .80s–.90s* published in the student ratings literature are usually *for the total or subscale collections of items*.

Item vs. total scale rating reliability. Despite the strong validity evidence, these differences in reliabilities raise a serious technical concern about the utility of global items, inasmuch as their ratings are used for summative personnel decisions about faculty. Their *potential for unacceptably low coefficients renders them inappropriate for any decisions about individual faculty members*. The strongest psychometric evidence rests with total and subscale ratings; the least stable is associated with the global items. Consider what foundation should be used to make decisions, for example, about contract renewal: One or two global items or the total scale rating based on 15 to 35 items.

Representativeness and Fairness

Item representativeness. After students have spent 45 hours, or a time close to that, in a course over the semester, suppose they rate the global item: "Overall, I rate this instructor as an excellent teacher," as part of a total scale. *Does the rating on that item seem to accurately capture the sum total of all of the teaching behaviors those students observed in their f2f or online course?* The percentages of explained variance in the research mentioned previously indicate it comes close. Can it represent that domain of behaviors? Will it reflect the differences between a f2f and online course?

There is no doubt that the item furnishes information about teaching effectiveness based on the validity studies, but should it be used for summative, super-important decisions about a professor's career? *Is one item rating of 0–3 or 4 an adequate, reasonable base from which teaching effectiveness can be inferred?* How fair is that? As Nuhfer (2010) argued, the evaluation of teaching, as a fractal form with complicated neural networks, is far too complex to be reduced to a single item.

Performance appraisal ratings. Hypothetically, could you accurately and fairly rate the performance of your

administrative assistant, department chair, or dean with one item to truly measure his or her degree of effectiveness? Would he or she want you to do that?

For more than half a century, *performance appraisals of employees in business and industry* typically have involved ratings by several professionals who are in the best positions to evaluate their performance (Bracken, Timmreck, & Church, 2001; Edwards & Ewen, 1996; Lepsinger & Lucia, 2009). These appraisals are significant because of the importance of the decisions and feedback to employees.

This evidence of job performance is based not only on the *ratings of all relevant job behaviors, but also by multiple qualified raters.* No single item can furnish that type of information. Evaluation of teaching effectiveness can and should be conducted similar to those performance appraisals with multiple ratings by different raters (Berk, 2009a, 2009b).

Professional Standards for Employee Decisions

Standards for Educational and Psychological Testing. One or two global ratings alone for major summative decisions about faculty performance are totally inadequate. That administrative practice *violates U.S. testing/scaling standards* according to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME Joint Committee on Standards, 1999). Clearly, these are *personnel decisions* about employees, not program decisions related to instruction or the curriculum. In the case of employee decisions like these, *one or two items do not reflect an accurate assessment of the instructor's job behaviors* (Standard 14.8). A total scale rating based on, for example, 35 items matched to the domain of teaching behaviors, or subscale ratings on specific areas of teaching competency would satisfy Standards 14.9 and 14.10 that require a *close link between the job content and the rating items.* Further, according to Standard 2.1, *reliability should be estimated for total, subscale, and combination ratings*, not for single items.

Personnel Evaluation Standards. The preceding standards are also supported by the latest edition of the *Personnel Evaluation Standards* (Joint Committee on Educational Evaluation Standards, 2009). The use of global items would not be upheld by Standards A4 (Valid Measurement) and A5 (Reliable Measurement). These standards require among other criteria that the scale include a representative sample of job tasks, inferences about the professor are drawn from the scale, and the ratings provide consistent (reliable) measurements of the professor's performance. *One item rating will rarely satisfy any of those requirements.*

Legal Standards for Employee Decisions

Employment instrument EEOC Guidelines. Since instruments can be used to hire, promote, demote, or fire an employee, the U.S. Equal Employment Opportunity Commission's (EEOC) *Uniform Guidelines on Employee Selection Procedures* set forth laws to protect an innocent employee from an evil employer who intentionally uses them to discriminate based on pay, age, color, disability, national origin, pregnancy, race, religion, sex, and sexual harassment (U.S EEOC, 2010). Such uses violate *federal anti-discrimination laws involving "disparate impact"* (practices that result in a disproportionate "adverse impact" on members of a minority group) *or "disparate treatment"* (practices that result in "intentional" discrimination of certain people groups during the hiring, promoting or placement process) (U.S EEOC, 2010).

Employment tests and other procedures, like *rating scales of job performance*, must be (1) "job-related and consistent with business necessity," and (2) "properly validated for the positions and purposes for which they are used" (U.S. EEOC, 2010). They cannot be designed, intended, or used to discriminate. Further, *employers are not permitted to adjust the scores, use different cutoff scores, or alter the results so as to discriminate against a particular group*. The complexity of this application of employment instruments and the "validation" requirements would *preclude global items from being used for any employment decisions about faculty.*

Employment instrument court cases. A long history of court cases on employment testing practices indicates that the *instrument used to measure employee performance must be based on a comprehensive job analysis of the job's tasks* related to a person's knowledge, skills, and abilities (KSAs) (Nathan & Cascio, 1986). In a review of 39 court of appeals cases and 43 district court cases from 2000–2007 (Ashe & U.S. EEOC, 2007), employment tests that produced "substantial adverse impact of a protected group" or "disparate impact" were scrutinized by the courts in terms of rigorous reliability and validity studies, especially in regard to selected cutoff scores for the decisions that resulted in the "impact." *One or two global items wouldn't come close to satisfying that level of scrutiny.*

Recommendations

Global items provide the illusion of (1) simplicity, (2) accurate and reliable information, and (3) the pinpoint precision needed for summative decisions about faculty. Unfortunately, the *single rating of a global item can be (a) unreliable, (b) unrepresentative of the domain of teaching behaviors it was intended to measure, and (c) inappropriate for personnel decisions according to U.S. professional and legal standards* (Berk, 2006). In fact, even if a global item satisfied the psychometric criteria, it still falls short in meeting professional and legal standards for employment decisions.

"Cease & Desist" Use of Global Items

Although administrators have used global items for decisions about faculty teaching performance for quite

some time and they are an attractive option, it is recommended that those practices come to a screeching halt. As noted previously, important, possibly career-changing, *individual personnel decisions are held to the highest standards psychometrically, professionally, and legally*, as they should be. If you know an administrator who is engaging in such practices, he or she should be urged to "*cease and desist*" before he or she is ordered legally to do so.

Four Alternatives to Global Item Rating

So what's an administrator supposed to do? Here are a few options to consider:

1. Use the *total scale rating* (mean/median) as the summary index across all items for the professor's courses over the past year. They can be displayed as simple numbers or graphically.
2. Use *two composite ratings*: one based on all items measuring *instructor* characteristics and a second based on those items measuring *course* characteristics. These ratings can also be presented for several courses.
3. Use *subscale (category) ratings* for different areas of teaching and course characteristics, which are consistent with the abundant evidence on the multidimensionality of student rating scales. Unfortunately, there is no agreement on the number of subscales based on the factors or dimensions that should be used for personnel decisions (Apodaca & Grad, 2005; Harrison et al., 2004; Hativa & Raviv, 1993; Renaud & Murray, 2005). Those factors will vary with different scales.

   The profile of subscale ratings furnishes information on the strengths and weaknesses of the professor and course. When administrators view these category ratings over time and courses, they can identify areas of growth and progress, or the "Peter Pan syndrome" (no growth). Some administrators may find that subscale ratings provide more information than they need. At least, all of these ratings are available.
4. Use *either of the preceding options in conjunction with data from other measures* of teaching effectiveness, such as self-ratings, teaching scholarship, administrator ratings, and teaching (course) portfolios.

Conclusion

These options would satisfy the *Standards for Educational and Psychological Testing* and *Personnel Evaluation Standards* cited previously. All of the preceding options, except 4, draw on ratings readily available from the student rating scale. It is simply a matter of how the ratings are reported for individual courses and across courses for each faculty member. Option 4 requires information from other sources, which would be preferable to using student ratings alone. Consider these options when reviewing report forms offered by vendors for online administration.

References

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness--generalizability of "N = 1" research: Comments on Marsh (1991). *Journal of Educational Psychology, 83*, 411-415.

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education) Joint Committee on Standards. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., Mohanty, G., & Spooner, F. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching, 52*(4), 134–141.

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education, 30*, 723–748.

Ashe, L., & U.S. EEOC. (2007, May). Employment testing and screening: Recent developments in scored test case law. Retrieved on August 20, 2012, from http://eeoc.gov/eeoc/meetings/archive/5-16-07/testcase_ashe.html.

Benton, S. L., Webster, R., Gross, A., & Pallett, W. (2010). *An analysis of IDEA Student Ratings of Instruction using paper versus online survey methods* (IDEA Technical Report No. 16). Manhattan, KS: The IDEA Center.

Beran, T., Violato, C., & Kline, D. (2007). What's the 'use' of student ratings of instruction for administrators? One university's experience. *Canadian Journal of Higher Education, 17*(1), 27–43.

Beran, T., Violato, C., Kline, D., & Frideres, J. (2005). The utility of student ratings of instruction for students, faculty, and administrators: A "consequential validity" study. *Canadian Journal of Higher Education, 35*(2), 49–70.

Berk, R. A. (2006). *Thirteen strategies to measure college teaching: A consumer's guide to rating scale construction, assessment, and decision making for faculty, administrators, and clinicians*. Sterling, VA: Stylus.

Berk, R. A. (2009a). Beyond student ratings: "A whole new world, a new fantastic point of view." *Essays on Teaching Excellence, 20*(1). (http://podnetwork.org/publications/teachingexcellence.htm.)

Berk, R. A. (2009b). Using the 360° multisource feedback model to evaluate teaching and professionalism. *Medical Teacher, 31*(12), 1073–1080. (DOI: 10.3109/01421590802572775)

Berk, R. A. (in press). Top 5 flashpoints in the assessment of teaching effectiveness. *Medical Teacher, 35*(1).

Bracken, D. W., Timmreck, C. W., & Church, A. H. (Eds.). (2001) *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes*. San Francisco: Jossey-Bass.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco: Jossey-Bass.

Cashin, W. E., & Downey, R. G. (1992). Using global student ratings for summative evaluation. *Journal of Educational Psychology, 84*, 563–572.

Cashin, W. E., Downey, R. G., & Sixbury, G. R. (1994). Global and specific ratings of teaching effectiveness and their relation to course objectives: Reply to Marsh (1994). *Journal of Educational Psychology, 86*, 649–657.

Centra, J.A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*, 1198–1208.

Edwards, M. R., & Ewen, A. J. (1996) *360° Feedback: The powerful new model*

for employee assessment and performance improvement. New York: American Management Association (AMACOM).

Ginns, P., & Barrie, S. (2004). Reliability of single-item ratings of quality in higher education: A replication. *Psychological Reports, 95*, 1023–1030.

Harrison, P. D., Douglas, D. K., & Burdsall, C. A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education*, *45*, 311–323.

Hativa , N., & Raviv, A. (1993). Using a single score for summative teacher evaluation by students. *Research in Higher Education*, *34*(5), 625–646. (DOI: 10.1007/BF00991923) (http://link.springer.com/10.1007/BF00991923)

Hoyt, D. P., & Lee, E-J. (2002, August). *Basic data for the revised IDEA system* (IDEA Technical Report No. 12). Manhattan, KS: Kansas State University Individual Development and Educational Assessment Center.

Joint Committee on Standards for Educational Evaluation. (2009). *The personnel evaluation standards: How to assess systems for evaluating educators* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Lepsinger, R., & Lucia, A. D. (2009). *The art and science of 360º feedback* (2nd ed.). San Francisco: Jossey-Bass.

Nathan, B. R., & Cascio, W. F. (1986). Introduction. Technical and legal standards. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 1–50). Baltimore, MD: Johns Hopkins University Press.

Otani, K., Kim, B. J., & Cho, J-Il. (2012). Student evaluation of teaching (SET) in higher education: How to use SET more effectively and efficiently in public affairs education. *Journal of Public Affairs Education, 18*(3), 531–544.

Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education*, *46*(8), 929–953. (DOI: 10.1007/s11162-005-6934-6)

U.S. Equal Employment Opportunity Commission (EEOC). (2010, September). Employment tests and selection procedures. Retrieved on August 20, 2012, from http://www.eeoc.gov/policy/docs/factemployment_procedures.html.

Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods, 4*(4), 361–375. (DOI: 10.1177/109442810144003)

Ronald A. Berk, Ph.D., is professor emeritus, biostatistics and measurement, and former assistant dean for teaching at The Johns Hopkins University. Now he is a full-time speaker, writer, and PowerPoint coach. He can be contacted at rberk1@jhu.edu, www.ronberk.com, or www.pptdoctor.net, and blogs at http://ronberk.blogspot.com.