Student and Instructor Use of the Teacher Behavior Checklist

Logan Stigall and Sarai Blincoe

Longwood University

Author Note

Logan Stigall, Department of Psychology, Longwood University; Sarai Blincoe, Department of Psychology, Longwood University.

The data analyzed for this study was collected by Sarai Blincoe while a doctoral student at the University of Kentucky.

Correspondence concerning this article should be addressed to Sarai Blincoe, Department of Psychology, Longwood University, 201 High Street, Farmville, VA 23909.

Email: blincoesc@longwood.edu

Abstract

Despite the demonstrated reliability and validity of the Teacher Behavior Checklist (TBC) as a tool for evaluating post-secondary instructors, research has yet to consider the impact of student, instructor, and course factors on TBC ratings. Additionally, it is unknown whether different evaluation sources converge when the TBC is used. This study examined self and student ratings of 35 instructors. The same characteristics which affect student ratings using other instrumentation impact student evaluation via the TBC. Further, several course and instructor variables impacted instructor self-evaluations. Instructor self-ratings did not correlate significantly with the ratings provided by the students. The implications of the findings for using the TBC for formative and summative assessment are discussed.

*Keywords*: Teacher Behavior Checklist, student evaluations, assessment, higher education

Student and Instructor Use of the Teacher Behavior Checklist

The evaluation of teaching has an important place in higher education. Summative assessments, including end-of-semester Student Evaluations of Teaching (SET; McKeachie, 1979), peer observations (Fernandez, 2007), and teaching portfolios (Berk, 2005) typically play a decisive role in promotion and tenure decisions. Formative assessments, based on student feedback, classroom visits by colleagues, and/or self-reflection, can help instructors make meaningful changes to classroom behavior in advance of formal summative evaluation (Penny, 2003). The SET, perhaps the most commonly used teaching assessment at the postsecondary level (Spooren, Brochx, & Mortelmans, 2013), typically evaluates such qualities as rapport, subject matter expertise, and overall teaching effectiveness, but varies considerably from institution to institution in length, item phrasing, and other elements of content and format (Spooren et al., 2013). Utilizing standardized and psychometrically sound instruments for both formative and summative assessments of teaching can strengthen the tenure and promotion process and facilitate instructor improvement.

A potential assessment tool, the Teacher Behavior Checklist, emerged from the efforts of Buskist, Sikorski, Buckley, and Saville (2002) to settle debate about what makes a master teacher. In the first part of their study, they asked undergraduate students to list three characteristics representative of a master teacher at the college level; a list of 47 characteristics resulted. A second group of undergraduates listed up to three behaviors corresponding with each of these characteristics. By comparing the lists of behaviors, the researchers combined the overlapping qualities, resulting in a final list of 28 items, which they called the Teacher Behavior Checklist (TBC; Buskist et al., 2002).

Since its development over 10 years ago, only a few studies have examined the psychometric properties of the TBC. When asked to rank the top ten qualities of a master teacher using the TBC, faculty and students agree on most of the items (Buskist et al., 2002). Schaeffer, Epting, Zinn, and Buskist (2003) replicated these findings and found similar agreement on the top qualities of a master teacher from faculty and students at a research university and a community college. In a study using student TBC ratings from the middle and end of the semester, Keeley, Smith, and Buskist (2006) found the expected test-retest reliability, given the tendency for evaluations to improve over the course of the semester. They also supported the strong internal reliability of the TBCs overall scale and two subscales, and the construct validity of the TBC through a comparison with their university’s standard course evaluation. When Keeley, Furr, and Buskist (2010) asked students to rate their best, worst, and most recent professor using the TBC they found that the instrument successfully differentiated among teachers. Thus evidence exists for the TBC’s reliability and validity, but research has yet to examine several questions that naturally arise for any measure potentially useful for the formative and summative assessment of instructors. The current study investigates three key questions which are outlined below.

**Research Question One: Which student and course qualities affect student ratings via the TBC?**

SETs were introduced into the classroom in the 1920s, but it was not until after 1969 that research on them rapidly increased (McKeachie, 1979). Much of the research sought to identify factors that influence student ratings of instructors. With regard to the TBC specifically, no research has yet explored the variables that influence student use of the instrument, though Keeley et al., (2006) specifically recommend looking at the effect of grade on TBC ratings. They express hope that the instrument is more resistant to these effects. This is the first study to assess the potential influence of a set of course characteristics and student characteristics on student ratings of instruction via the TBC.

Consistent with findings for other SET forms, we predict that higher student ratings on the TBC will correlate positively with instructor experience (McPherson, Jewel & Kim, 2009; McPherson & Jewell, 2007) and negatively with class size (Bedard & Kuhn, 2008; McPherson, 2006). Students in higher level courses (Santhanam & Hicks, 2002), and those expecting a higher grade (Marsh & Roche, 1997; Spooren et al., 2013), are predicted to give more positive ratings. The influence of other factors, such as instructor and student sex, and student effort, are more difficult to anticipate. Female students sometimes give higher SET ratings than male students (Kohn & Hartfield, 2006; Santhanam & Hicks, 2002); Basow, Phelan, and Capotosto (2006) found this pattern only for female instructors, but Kohn and Hartfield (2006) found this pattern only for male instructors. Thus, we hypothesize that female students will give more positive ratings than males, but explore the possibility that instructor sex moderates this relationship. With regards to student effort, Marsh and Roche (2000) found a positive relationship with course workload but Centra (2003) found that courses in which students feel difficulty, workload, and pace are “about right” received the highest ratings. We predict that effort, operationalized as number of hours worked outside of class, will positively correlate with student ratings. Finally, we examine the possible influence of instructor teaching style (an emphasis on lecture vs. discussion) by asking whether the number of minutes the instructor speaks in a typical class, relates to student ratings.

**Research Question Two: Which course and instructor qualities affect self-ratings via the TBC?**

Another form of evaluation is instructor self-assessment. Portfolios, which often include teaching goals, student ratings, and work samples, are a common form of self-reflection (Berk, 2005). Although portfolios are a potentially useful tool for teaching improvement, instructors rarely use them to full potential because they require considerable work to assemble, regular updating, and may not be valued in the summative (i.e., promotion and tenure) process (Cerbin, 1994). Given the benefits of an instructor’s formative self-assessment for the improvement of teaching, and in the face of the considerable drawbacks of portfolio construction, the TBC may serve as a helpful tool. The TBC items allow faculty to reflect on specific characteristics and behaviors, thus focusing improvement efforts, and the survey format makes for a drastically reduced time commitment. It also has the potential for use in conjunction with larger portfolio projects. Unfortunately, no studies on the TBC have considered how instructors use the measure, although Schaeffer et al. (2003) do report that faculty members endorsed similar TBC items as important for college level master teaching, regardless of gender and appointment (adjunct vs. full-time).

The larger literature on evaluations of teaching tends to neglect self-assessment, or focus on best practices and qualitative discussion to the neglect of quantitative explorations (Berk, 2005; Cerbin, 1994). The current study is a first step to filling this considerable gap in the literature on teaching assessment. Based largely on the research on student evaluations of teaching, we predict higher self-ratings on the TBC will correlate positively with instructor experience. Additionally, we hypothesize that class size will correlate negatively with TBC ratings, perhaps because instructors feel more comfortable in a smaller class setting. Instructors teaching higher level courses are predicted to give higher self-ratings compared to instructors teaching lower level courses, but we do not expect instructor sex to influence instructor self-ratings (Schaeffer et al., 2003). Finally, we examine the possibility of a correlation between self-ratings and the time an instructor spends speaking during a typical class session.

**Research Question Three: Do evaluation sources tell the same story?**

A multidimensional approach to teacher evaluation is generally preferred because multiple sources provide a more accurate and reliable picture (Ackerman, Gross & Vigernon, 2009; Berk, 2005; Spooren et al., 2013). Sources of evaluation include student evaluations, self-reflection, and observation by peers, administrators, and external reviewers (Feldman, 1989). To be maximally useful, these multiple sources should provide a converging picture of instructional quality. Feldman (1989) conducted a meta-analysis of studies that reported correlations between sources. Most frequently, studies compared the evaluations of current students with self (*N* = 19; *r* = .29), colleague (*N* = 14; *r* = .55), and administrator (*N* = 11; *r* = .39) evaluations. If the TBC is a psychologically sound instrument, we would expect to find similar or stronger correlations between sources. Although past research has examined the correspondence of evaluations, this is the first study to compare self-ratings with student ratings using the TBC. Based on past research, we expect that a small (but statistically significant) positive relationship between self-ratings and ratings provided by current students.

**Method**

**Participants**

**Instructors**. During the 2010-2011 academic year at a large Midwestern university, 35 (16 females, 19 males, 88.5% Caucasian) instructors volunteered to participate in a larger study of student incivility. Instructor disciplines included fine arts/humanities (*n* = 8; e.g., philosophy, English, foreign languages), social/behavioral sciences (*n* = 10; e.g., psychology, sociology, political science), physical sciences (*n* = 11; e.g., chemistry, astronomy, mathematics, physics) and miscellaneous courses (*n* = 6; e.g., first year seminars, accounting). At the time of the study, the majority of instructors (*n* = 26; 74%) possessed a doctorate degree (or comparable terminal degree); six held the rank of full professor, nine held the rank of associate professor, six held the rank of assistant professor, and 14 held the rank of instructor or non-tenure track.

**Students.** Each instructor designated one of their courses for participation in the larger study. Fourteen of the courses were at the 100 level, 11 at the 200 level, and 10 at the level of 300 or above. Between 14 and 183 students (*M* = 49.45; *SD* = 39.75) participated from each of the 35 classes enrolled in the project, for a total of 1,632 students (56.5% females, 43.5% males). The student sample was primarily Caucasian (87.5%) but also included students who identified as African American (6.5%), Asian American (3.0%) and other (3.0%). The class standings of the students included freshmen (30.4%), sophomores (29.3%), juniors (22.9%), seniors (16.0%), and other standing (1.4%).

**Materials**

**Student background survey.** Students reported demographic information (i.e., biological sex, race, age, class standing) along with course information, including expected grade and effort (hours spent on class outside of class time). Students rated the instructor on three general quality items using a 1 (*not at all*) to 7 (*definitely*) scale(i.e., I would take another course with this instructor; I wish I hadn’t signed up for this class–reverse coded; I would recommend this instructor to a friend) which were combined to form a mean Instructor Quality variable (α = .80).

**Instructor background survey.** Instructors reported demographic information (i.e., biological sex, race) along with educational background (i.e., highest degree attained and year of attainment). Next, they answered questions about employment, including their academic discipline, rank, and length of employment at the current university. Finally, they reported how many minutes they spent speaking, on average (e.g., lecturing, leading discussion, answering questions), during a typical 50 minute class period (i.e., Time Speaking), along with other questions about teaching style not analyzed for the present study.

**Teacher Behavior Checklist.** The TBC uses a five-point scale ranging from A (*Always exhibits/has exhibited behaviors reflective of this quality*) to E (*Never exhibits/has exhibited behaviors reflective of this quality*). Instructions for students prompted them to rate the instructor of the course in which they currently sat. Instructions for instructors prompted them to rate themselves while keeping in mind the specific course that was observed for the study.

All 28 TBC qualities load on one factor with item factor loadings ranging from .47 to .79 and excellent inter-item reliability (Keeley et al., 2006). An acceptable two factor solution loads 13 items on a *caring and supportive* subscale and 11 items on a *professional competency and communication skills* subscale, with four items not used (Keeley et al., 2006). Henceforth, we refer to these subscales as Caring and Competency, respectively. We followed recommendations from Keeley et al. (2006) to utilize the total score and the two subscale scores. To calculate the mean for all three scales, we scored As as 5s to Es as 1s, where higher scores indicated more positive ratings. Consistent with past research, we found high internal reliability for all student rating scales and acceptable internal reliability for the instructors on both the Total and Caring scales (see Table 1). The Instructor Competency subscale rating, however, failed to meet standard acceptable reliability (see Table 1), which may influence the results.

**Procedure**

Instructors were recruited via email, phone, and face-to-face contact, for participation in a university-wide study of student incivility and its relation to instructor and course variables. Participating instructors consented to have an undergraduate research assistant visit one regularly scheduled class session. Fall 2010 visits occurred between October 28th and December 10th and Spring 2011 visits occurred between April 1st and April 15th.

At the class visit, the research assistant videotaped a three minute portion of the lecture at the beginning of the class and at the end of the visit rated the instructor on several scales.1 At the end of the visit, the students in attendance were invited to participate by completing a questionnaire packet containing the student background survey, Student TBC, and other questionnaires not analyzed for the present study. Participating students did not receive any form of compensation.

Also at the end of the class session, in a different location, the instructor completed a questionnaire packet containing the instructor background survey, Instructor TBC, and other questionnaires not analyzed for the present study. Additionally, at the conclusion of the semester, participating instructors provided a course syllabus and grade distribution and received $20 for their participation in the study. Information about the course level and class size was retrieved from the syllabus and online sources.

**Results**

**Data Preparation and Preliminary Analyses**

Three TBC self-rating scores (Competency subscale, Caring subscale, Total score) were calculated for each instructor, except for the two instructors that did not complete the TBC measure. These Instructor TBC ratings are used in the examination of Research Questions Two and Three.

If a student failed to rate an item on the TBC, that missing rating was replaced with the mean item rating from the class, rounded to the nearest whole number, for up to four missing items; less than .05% of the items were replaced. Students missing responses for more than four items were excluded from the analyses (*n* = 79 students; 4.58%). We also excluded students who demonstrated a clear response bias as indicated by identical ratings for all TBC items and the 15 item questionnaire which preceded the TBC in the student packet (*n* = 9 students; 0.5%). If a student responded to a TBC item with more than one rating (A/B), the lower rating (B) was used (*n* = 12 items, < .01%). Three TBC scores (Competency subscale, Caring subscale, Total scale) were calculated for each student. For a given class, we took the Student TBC ratings and averaged them. Thus we had, for each of the 35 classes, three Class TBC scores (Competency subscale, Caring subscale, Total Score). The Student TBC ratings are used to examine Research Question One. The Class TBC scores are used to examine Research Questions Two and Three.

For the student effort variable, we excluded from the analyses any student that reported above 15 hours a week based on an outlier calculation of three standard deviations; only 1% of students reported effort of greater than 15 hours/week.

Table 1 reports the means and standard deviations for the three Instructor TBC ratings, the three Class TBC ratings, as well as the continuous variables of class size (Class Size), instructor experience (I Years Experience), estimated instructor speaking time (I Speaking Time), and student effort (S Effort). All three Class TBC ratings correlated highly with one another, as did all three Instructor TBC ratings (see Table 1).

Each analysis that involved the TBC was conducted using all three scales (Total, Caring, Competency). Where results are reported for the Total scale only, the reader can assume that the Caring and Competency subscales performed identically, unless indicated otherwise.

**Research Question One: Student Ratings via the TBC**

Three factorial ANOVAs tested the effect of student and instructor biological sex on Student TBC ratings. For the Total score, there was a significant main effect of instructor sex, *F*(1, 1624) = 36.49, *p* < .001, η2 = .022, but student sex did not have a significant main effect, *p* = .38. A significant interaction, *F*(1, 1624) = 9.49, *p* = .002, η2 = .006, qualified the main effect of instructor sex. When rating female instructors, male students gave significantly lower ratings (*M* = 4.31, *SD* = .55) than females (*M* = 4.43, *SD* = .54), *t*(801) = 3.20, *p* = .001, *r* = .11. When rating male instructors, male (*M* = 4.21, *SD* = .69) and female (*M* = 4.14, *SD* = .72) students did not differ in their ratings, *p* = .158. Thus, our prediction regarding student sex was supported, though a significant interaction qualified the main effect. Biological sex of the student and the instructor interacted to influence Student TBC ratings. The highest ratings went to female instructors when female students did the rating.

In support of our hypothesis regarding the influence of grade on student ratings, a one-way ANOVA revealed a significant main effect of grade on evaluation of instructor, *F*(2, 1606) = 55.81, *p* < .001, η2 = .065. Tukey post hoc tests showed that students who expected an A (*M* = 4.46, *SD* = .51) rated their instructor significantly higher than students who expected a B (*M* = 4.18, *SD* = .66), who in turn gave significantly higher ratings than students who expected a C or worse (*M* = 4.03, *SD* = .78), *p*s < .001. A correlation was used to test the prediction that student effort would associate positively with student ratings (see Table 1). Student effort and Student TBC ratings were significantly related, but contrary to our prediction, as students reported spending more time on the class they gave lower TBC ratings.

A one-way ANOVA revealed a trend for course level impacting Class TBC ratings on the Total scale *F*(2, 32) = 2.54, *p* = .095, η2 = .14. In partial support of our hypothesis, students in the 300+ level courses did give significantly higher mean ratings (*M* = 4.52; *SD* = .22) than those in the 200 level courses (*M* = 4.20; *SD* = .27), *p* = .035, and there was a trend toward a significant difference between ratings in the 300+ level courses and the 100 level courses (*M* = 4.30; *SD* = .41), *p* = .114. Course level had no effect on the Class ratings of Caring and Competency.

A sizeable non-significant relationship of *r*(33) = -.30, *p* = .09 was found between class size and Class TBC Caring, suggesting that as class size increases, perceptions of instructor caring may decrease. Class size did not significantly correlate with Class ratings on the Total scale or Competency subscale (See Table 1). These results provide partial support for our prediction, such that smaller classes tend to give higher ratings to instructors, but only on the Caring subscale. As with class size, instructor self-reported average Time Speaking (ranging from 20-50 minutes; *M* = 36.56, *SD* = 10.09) correlated significantly only with the Caring subscale. As an instructor reports spending more time speaking in class, Class ratings of instructor caring decrease. Analyses failed to support our prediction for the relationship between instructor experience and Class TBC ratings; no relationship existed (see Table 1).

**Research Question Two: Instructor Self-Ratings via the TBC**

A series of analyses tested the hypotheses regarding the possible influence of instructor and course characteristics on instructor use of the TBC for self-evaluation. Regarding instructor characteristics, three independent sample *t-*tests supported our prediction that self-ratings would not differ by instructor sex; male and female instructors gave similar self-ratings on all three TBC scales, *p*s > .41.

Instructor experience did not significantly correlate with self-ratings (see Table 1) but an examination of the scatterplots suggested a curvilinear relationship as a source of distortion in the correlation coefficient. Using a curve estimation regression procedure, the Instructor Total TBC was predicted from experience. Neither the linear [*F*(1, 31) = .525, *p* = .474] nor quadratic [*F*(2, 30) = 2.85, *p* = .073] equations reached significance, but the cubic equation was significant, *F*(3, 29) = 7.144, *p* = .001, *R*2 = .425), partially supporting our prediction. TBC ratings appear to be highest among those most recently graduated, with ratings decreasing until around 7 years of experience. Another increase in ratings then occurs, up until about 25 years of experience, with a peak comparable to that of the early career. This is followed by another decline. Notably, this relationship may be driven by the qualities of this sample, which overrepresented those in early and late career stages, with a noticeable absence of those between 10 and 20 years post-degree.

As for the relationships between course characteristics and instructor self-ratings, course level had no effect on instructor self-ratings (*ps* = .162 [Total]; .118 [Caring]; .409 [Competency]). Class size correlated significantly only with the Caring subscale (see Table 1). As an instructor’s class size increased, self-ratings of caring decreased, partially supporting our hypothesis. The instructor’s estimation of their Time Speaking during a class session significantly correlated with TBC ratings, but only for the Competency subscale (see Table 1). As the instructor’s estimate of the typical amount of time they speak in a class increased, self-ratings of competency also increased.

**Research Question Three: Source Convergence on the TBC**

Correlational analyses were conducted to test our hypothesis concerning the associations between the Instructor and Class TBC ratings. Instructor self-ratings were not significantly correlated with the Class ratings for any of the TBC scales;however, for two of the six associations, the size of the correlation coefficient is comparable to that reported in the Feldman (1989) meta-analysis (see Table 1). Therefore, the failure to reach statistical significance may be due to insufficient power in this analysis. Importantly for the construct validity of the TBC, the Instructor Quality measure, based on three items rated by students, also correlated significantly with the Class TBC Total score, *r*(1593) = .68, *p* < .001.

**Discussion**

The Teacher Behavior Checklist is a promising instrument that may prove useful for both formative and summative assessments of instructor quality. The easy-to-use checklist format incorporates behavioral anchors, allowing instructors to tailor their improvement efforts to the acquisition of specific characteristics associated with master teaching (Schaeffer et al., 2006). To date, multiple studies support the internal reliability and construct validity of the TBC (Keeley et al., 2006; Keeley, English, Irons & Henslee, 2013), but the effects of course, student, and instructor characteristics on TBC use remained unexplored. This study was the first to investigate the ways in which student (i.e., biological sex, effort, expected grade), course (i.e., class size, class level) and instructor (i.e., biological sex, experience, time speaking) characteristics might influence or associate with student ratings and self-ratings via the TBC. Furthermore, this study provides preliminary evidence as to how different evaluators (student and self) use the TBC such that ratings of instructor quality may converge or diverge across different sources. Gathering ratings from multiple sources is a useful way to understand what changes to make. The instructor self-ratings did not significantly correspond with the student ratings. These results suggest that the TBC is a useful tool for evaluating instructors but should be used by many sources to get a more accurate picture.

In their suggestions for future research, Keeley et al., (2006) expressed hope that the behavioral anchors of the TBC would reduce the effects of such variables as student grade on teacher evaluation. Our analyses suggest that despite its checklist format, behavioral anchors, and foundation in shared impressions of what makes a master teacher, use of the TBC to evaluate instructors is affected by the same characteristics that influence other assessment tools (Spooren et al., 2013).

Higher student ratings of instructors were associated with female instructors, higher expected grades, upper-level courses, and smaller classes. Contrary to previous research (McPherson et al., 2009; McPherson & Toff Jewell, 2007), instructors with more experience did not receive higher student ratings compared to instructors with less experience. The distribution of experience levels in this particular sample may account for this discrepancy. Alternatively, we used number of years since degree as a proxy for teaching experience; the two are likely correlated, but not necessarily identical.The negative correlation of student ratings with student effort runs counter to the one other study that found higher student evaluations of teaching for teachers who encouraged more effort from students (Heckert, Latier, Ringwald-Burton, & Drazen, 2006). Operationalization of effort may account for the differences, and additional research on this variable is clearly needed.

The TBC Total scale performed identically to the subscales for all student characteristics (i.e., student sex, expected grade, effort). Interestingly, the results for the Total scale diverged from those for the Caring subscale on the class size and instructor speaking time variables, both of which negatively correlated with Caring. A larger class size poses challenges to an instructor’s ability to cultivate personal relationships with each student, reflected by such Caring subscale items as *accessible*, *rapport*, and *understanding*. Instructors may tend to rely more on lecture in larger classes, thus reducing scores for such Caring scale qualities as *promotes class discussion* and *provides constructive feedback*.

Instructor sex, student sex, expected grades, student effort, course level, class size, and instructor speaking time were all associated significantly with student usage of the TBC. Knowing that, for example, higher expected grades and lower expended effort are associated with more favorable student ratings, untenured or non-tenure-track faculty may be tempted to relax course standards in the hopes of inflating course evaluations. Our findings indicate that the TBC may not perform better than other measures if used for summative assessment with student raters. When it comes to promotion decisions, an understanding of the factors that influence ratings is essential. For example, if one professor is only teaching large introductory courses while another professor is teaching small, upper-level courses, the latter might receive higher ratings even though they might not necessarily be a better teacher.

As a self-evaluation tool, the TBC proved relatively resistant to the effects of instructor sex and course level. The variables of class size and instructor speaking time correlated only with ratings on the most relevant subscales (Caring and Competency, respectively), perhaps providing evidence for construct validity rather than simply biased assessment. The patterns found for class size replicate those found for the student ratings. Both instructors and students perceive the instructor teaching a smaller class as higher on the Caring subscale. Interestingly, the pattern for the instructor speaking time diverged from that found for the student ratings. Whereas class ratings of caring decreased as instructor speaking time increased, instructor self-ratings of competency actually increased as speaking time increased. Perhaps from the instructor’s perspective, speaking more during class is associated with higher ratings on qualities such as *authoritative, effective communicator, knowledgeable about subject,* and *prepared*.2 Future studies should expand on the student, instructor and course qualities identified here as affecting TBC ratings (see Spooren et al., 2013 for additional characteristics known to impact SET validity).

With regard to the third research question, the instructor self-ratings did not correlate with the student ratings at the traditional *p* = .05 significance level. A focus on the significance level suggests that these findings conflict with past research that found a small to medium average association between self-ratings and student ratings (Feldman, 1989). However, given the moderate size of the correlation coefficients obtained in this study, an association may exist that our small sample was underpowered to detect with traditional significance testing. Interestingly– given that the only significant association with instructor ratings was with students on the Caring scale– Schaeffer et al. (2003), using the TBC, found that students placed greater value on the student-teacher relationship, including Caring scale qualities, such as *rapport* and *accessible,* than faculty. The lack of correspondence for the Competency scale may be due to the low internal reliability of that scale (discussed further below).

**Limitations and Future Directions**

Certain elements of the study design limit the conclusions that can be drawn from the results. Several factors likely hindered the process of recruiting instructors for the larger study of teaching and learning from which these data derive, including the research emphasis of the university at which the study was conducted, the time and coordination required for participation, and the reluctance to invite others into the classroom (often considered a sacred and highly independent space) for purposes of observation and evaluation. As a result of these factors, as well as the limited resources of the study team, the number of instructors participating in the study was limited to 35. The sample size limits the power of all analyses that included instructor self-ratings and class-level mean ratings. Despite this issue, significant associations and differences were obtained for all variables (suggesting that the associations are not spurious) with the notable exception of course level, which had only 10-15 classes per category. Future research would benefit from additional study of the effects of course level by obtaining larger numbers of participating instructors or by requesting that instructors provide ratings for more than one course. In this way, comparisons can be made between an instructor’s self ratings for introductory courses as opposed to upper-level courses.

In addition to limited power, the reliability for the instructor Competency subscale did not meet an acceptable level. This problem is likely attributable to the small sample. It is important to note that the reliability of these measures is only important in situations that utilize overall ratings, such as the typical tenure and promotion scenario. If the TBC is used to identify specific areas for improvement, with a focus on individual items rather than overall scores, the possibility of an internal reliability issue for self-ratings can be side-stepped.

In conducting the analyses of correspondence between sources, and the effects of course characteristics on student ratings, a mean Class TBC rating was calculated. The use of this value may limit our conclusions. Although other studies of the correspondence between different evaluation sources rely on mean SET ratings (see for example, Feldman’s 1989 meta-analysis), this approach sacrifices variability and gives the same weight to a mean SET based on 20 students as it does to one based on 200 students. Notably, the flaws of this approach apply to the use of SETs more broadly. Evaluations of teaching, including those for promotion and tenure, typically rely on mean SETs and comparisons of those means to department, college, and/or university means.

Future studies of the TBC should incorporate observer ratings in a procedure that closely mimics that of the peer review process by incorporating multiple observers, multiple class visits, and granting access to supplementary materials such as course syllabi. Perhaps more importantly, additional research should explore how instructors might incorporate the TBC into a formative review process. Perhaps by completing the TBC themselves (or having students or colleagues complete it) at the beginning, midpoint, and end of a semester, instructors can more tailor their efforts to improve instructional quality. The behavioral anchors of the measure could facilitate improvement by giving instructors specific behaviors to refer to and implement.

**Conclusions**

The Teacher Behavior Checklist offers instructors and administrators an alternative approach to both formative and summative assessment of teaching. Our preliminary findings suggest that with appropriate awareness of the characteristics that impact its use, the TBC may serve as a psychometrically sound instrument for assessment of instructors. In particular, it has potential to work as a valuable self-assessment tool, in combination with, or as an alternative to, the more time-intensive teaching portfolio.

References

Ackerman, D., Gross, B. L., & Vigneron, F. (2009). Peer observation reports and student evaluations of

teaching: Who are the experts? *Alberta Journal of Educational Research*, *55*(1), 18-39.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their

best and worst professors. *Psychology of Women Quarterly*, *30*(1), 25-35. doi: 10.1111/j.1471-6402.2006.00259.x

Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of

instructor effectiveness. *Economics of Education Review*, *27*(3), 253-265. doi: 10.1016/j.econedurev.2006.08.007

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of*

*Teaching and Learning in Higher Education*, *17*(1), 48-62.

Buskist, W., Sikorski, J., Buckley, T., & Saville, B. K. (2002). Elements of master teaching. In S. F.

Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer* (pp. 27–39). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less

course work? *Research in Higher Education*, *44*(5), 495-518. doi: 10.1023/A:1025492407752

Cerbin, W. (1994). The course portfolio as a tool for continuous improvement of teaching and

learning. *Journal on Excellence in College Teaching*, *5*(1), 95-105.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves,

current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education*, *30*(2), 137-194. doi: 10.1007/BF00992716

Fernandez, C. E., & Yu, J. (2007). Peer review of teaching. *The Journal of Chiropractic Education*,

*21*(2), 154.

Heckert, T. M., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort,

perceived class difficulty appropriateness, and student evaluations of teaching: Is it possible to" buy" better evaluations through lenient grading? *College Student Journal*, *40*, 588-596.

Keeley, J. W., English, T., Irons, J., & Henslee, A. M. (2013). Investigating Halo and Ceiling Effects in

Student Evaluations of Instruction. *Educational and Psychological Measurement*, *73*(3), 440-457. doi: 10.1177/0013164412475300

Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating psychology students' perceptions of

teachers using the teacher behavior checklist. *Teaching of Psychology*, *37*(1), 16-20. doi: 10.1080/00986280903426282

Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its

utility for evaluating teaching. *Teaching of Psychology*, *33*(2), 84-91. doi: 10.1207/s15328023top3302\_1

Kohn, J., & Hatfield, L. (2006). The role of gender in teaching effectiveness ratings of faculty. *Academy*

*of Educational Leadership Journal*, *10*(3), 121-137.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective:

The critical issues of validity, bias, and utility. *American Psychologist*, *52*(11), 1187. doi: 10.1037/0003-066X.52.11.1187

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students'

evaluations of teaching: popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, *92*(1), 202. doi: 10.1037/0022-0663.92.1.202

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe*, *65*(6), 384-397. doi:

10.2307/40248725

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *The Journal of Economic*

*Education*, *37*(1), 3-20. doi: 10.3200/JECE.37.1.3-20

McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores

be adjusted? *Social Science Quarterly*, *88*(3), 868-881. doi: 10.1111/j.1540-6237.2007.00487.x

McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What determines student evaluation scores? A

random effects analysis of undergraduate economics classes. *Eastern Economic Journal*, *35*(1), 37-51. doi: 10.1057/palgrave.eej.9050042

Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching:

Four shortcomings of SRT research. *Teaching in Higher Education*, *8*(3), 399-411. doi: 10.1080/13562510309396

Santhanam, E., & Hicks, O. (2002). Disciplinary, gender and course year influences on student

perceptions of teaching: Explorations and implications. *Teaching in Higher Education*, *7*(1), 17-31. doi: 10.1080/13562510120100364

Schaeffer, G., Epting, K., Zinn, T., & Buskist, W. (2003). Student and faculty perceptions of effective

teaching: A successful replication. *Teaching of Psychology*, *30*(2), 133-136.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching

The state of the art. *Review of Educational Research*, *83*(4), 598-642. doi: 10.3102/0034654313496870

Footnotes

1Five different research assistants participated in this study. One of the scales completed by the research assistant was the Teacher Behavior Checklist. This scale was included with the intention of obtaining an “outsider observer” perspective on the instructor’s teaching. Correlational analyses found that for all three TBC scales, observer ratings significantly correlated, in the positive direction, with the student ratings (all *rs* > .44). There were no significant correlations between the observer and the instructor perspectives. As such, our observer perspective appears to represent an additional student perspective. The observer ratings were further limited by the internal reliability of the Competency subscale (α = .45).

2During the classroom visit, the research assistant in attendance used a stopwatch to record the total amount of time the instructor spoke during the class session. Ultimately, we had speaking times for 31 of the 35 classes. The instructor self-reported speaking time correlated positively, but not significantly, with this observation-based measure, *r*(30) = .33, *p* = .07. The observer records of speaking time are likely to be affected by the practical challenges of keeping track of speaking time during a live class, while making other observations. Ultimately, we found the same pattern of significant correlations using the instructor self-reported speaking time as the observation-based measure; namely, a negative correlation with class ratings of caring and a positive correlation with instructor self-rating of competency.

Table 1. Means, Standard Deviations, and Correlation Coefficients for TBC Ratings and Continuous Variables

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1. S Effort | – |  |  |  |  |  |  |  |  |  |
| 2. Class size |  | – |  |  |  |  |  |  |  |  |
| 3. I Years Experience |  |  | – |  |  |  |  |  |  |  |
| 4. I Time Speaking |  |  |  | – |  |  |  |  |  |  |
| 5. S Total | -.15\* | -.21 | .14 | -.21 | – |  |  |  |  |  |
| 6. S Caring | -.16\* | -.30† | .15 | -.29† | .97\* | – |  |  |  |  |
| 7. S Competence | -.14\* | -.09 | .11 | -.07 | .95\* | .85\* | – |  |  |  |
| 8. I Total |  | -.22 | -.13 | .10 | .11 | .17 | .00 | – |  |  |
| 9. I Caring |  | -.43\* | .03 | -.15 | .30† | .30 | .09 | .87\* | – |  |
| 10. I Competence |  | .08 | -.25 | .38\* | .01 | .05 | .01 | .80\* | .44\* | – |
| *Mean* | 3.79 | 49.45 | 15.54 | 36.56 | 4.33 | 4.26 | 4.41 | 4.31 | 4.25 | 4.40 |
| *SD* | 2.90 | 39.75 | 12.71 | 10.09 | .34 | .40 | .31 | .27 | .34 | .28 |
| *Cronbach’s α* | – | – | – | – | .96 | .92 | .89 | .80 | .75 | .55 |

\* *p* < .05, † *p* < .10, “I” = Instructor, “S” = Student