TOPICAL ARTICLES

The Teacher Behaviors Checklist: Factor Analysis of Its Utility for Evaluating Teaching

Jared Keeley, Dale Smith, and William Buskist
Auburn University

We converted the Teacher Behaviors Checklist (TBC; Buskist, Sikorski, Buckley, & Saville, 2002) to an evaluative instrument to assess teaching by adding specific instructions and a Likert-type scale. Factor analysis of the modified TBC produced 2 subscales: caring and supportive and professional competency and communication skills. Further psychometric analysis suggested the instrument possessed excellent construct validity and reliability, underscoring its potential as a tool for assessing teaching. This instrument clearly identifies specific target teaching behaviors that instructors can alter to attempt to improve their teaching effectiveness.

Research on master teachers, especially studies aimed at uncovering the determinants of master teaching, traditionally focus on personality traits or qualities of teachers known for excellence. This research entails studies of award-winning teachers (e.g., Baiocco & DeWaters, 1998), analyses of student evaluations (e.g., Feldman, 1976), and empirical studies (e.g., Epting, Zinn, Buskist, & Buskist, 2004) and is accented by the reflective writing of outstanding teachers (e.g., Brookfield, 1990; Palmer, 1998). Among other things, this research portrays master teachers as passionate, respectful, approachable, creative, fair, understanding, and well prepared (for a review, see Buskist, Sikorski, Buckley, & Saville, 2002).

In an attempt to provide behavioral anchors for the typical global personality descriptors that permeate this literature, Buskist et al. (2002) developed the Teacher Behaviors Checklist (TBC), a 28-item inventory that defines such personality qualities in terms of the behaviors that comprise them. Using undergraduate participants, they asked students to list the qualities of effective teachers, defined as teachers from whom students had learned a great deal and who made the learning process enjoyable. These researchers catalogued the personality characteristics provided by the students and then asked another group of students to identify specific behaviors reflective of these characteristics. To determine the most important qualities and behaviors of these teachers, Buskist et al. asked undergraduates and faculty at a doctoral-level institution to rate the top 10 qualities and behaviors from among the 28 items. Students and faculty showed strong agreement in their ratings, a finding replicated at both the community college (Schaeffer, Epting, Zinn, & Buskist, 2003) and baccalaureate levels (Wann, 2001).

Based on the success of the TBC in identifying undergraduate and faculty perspectives on the qualities and corresponding behaviors of effective teachers, we examined the TBC's utility as an instrument for assessing teaching. Because the TBC is a behaviorally based scale, its potential for formative evaluative purposes is of particular interest to instructors wishing to improve their teaching. If a teacher receives poor ratings on a specific characteristic, he or she can attempt to make improvements in behaviors reflective of that characteristic.

Study 1

The first study evaluated the basic psychometric properties of the TBC through an exploratory factor analysis, examination of internal reliability of the scales, and a comparison with the standard Auburn University teaching evaluation. This approach allowed us to examine the basic factor structure of the instrument as well as measures of its construct validity and internal reliability.

Method

Participants. We recruited 313 Auburn University students enrolled in four introductory psychology sections (*n*s = 86, 130, 53, 44) to participate in this study. A different professor taught each section (three full-time faculty and one part-time instructor). Students ranged in age from 19 to 45 years ($M = 19.9$ years, $SD = 2.2$).

Materials. We converted the TBC to an evaluative inventory (see Table 1) by adding a set of instructions and a 5-point Likert-type scale ranging from 5 (*frequent*; indicated by A) to 1 (*never*; indicated by E), with a midpoint of 3 (*I have no opinion*). Students rated their specific professor by filling in a computer scoring sheet. Students also wrote their age on this sheet.

Students also completed a standard Auburn University end-of-semester teaching evaluation. This evaluation con-

Table 1. Variations of the Teacher Behaviors Checklist Used in Studies 1 and 2

Instructions used in Studies 1 and 2
Instructor Evaluation—Dr. _____
Instructions: On the back of this sheet are 28 teacher qualities and the behaviors that define them. Please rate Dr. _____ on the extent to which you believe she (or he) possesses these qualities and exhibits the corresponding behaviors.
Please use the following scale for your ratings by bubbling in the corresponding space in your scantron for each question/item number.
Likert scale used in Study 1
A = Dr. _____ frequently exhibits/has exhibited these behaviors reflective of this quality.
B = Dr. _____ sometimes exhibits/has exhibited these behaviors reflective of this quality.
C = I have no opinion on whether Dr. _____ exhibits/has exhibited these behaviors reflective of this quality.
D = Dr. _____ rarely exhibits/has exhibited these behaviors reflective of this quality.
E = Dr. _____ never exhibits/has exhibited these behaviors reflective of this quality.
Likert scale used in Study 2
A = Dr. _____ always exhibits/has exhibited these behaviors reflective of this quality.
B = Dr. _____ frequently exhibits/has exhibited these behaviors reflective of this quality.
C = Dr. _____ sometimes exhibits these behaviors reflective of this quality.
D = Dr. _____ rarely exhibits/has exhibited these behaviors reflective of this quality.
E = Dr. _____ never exhibits/has exhibited these behaviors reflective of this quality.

Instructions for providing written commentary used in Studies 1 and 2
In addition, please use the space below on this side of the page to write any comments regarding Dr. _____ teaching. These comments may include both what you find positive and negative about Dr. _____ teaching.
Please be sure to read each item in this list carefully. Thank you.

Item	Teacher Qualities and Corresponding Behaviors
1	*Accessible* (Posts office hours, gives out phone number and e-mail information)
2	*Approachable/Personable* (Smiles, greets students, initiates conversations, invites questions, responds respectfully to student comments)
3	*Authoritative* (Establishes clear course rules; maintains classroom order; speaks in a loud, strong voice)
4	*Confident* (Speaks clearly, makes eye contact, and answers questions correctly)
5	*Creative and Interesting* (Experiments with teaching methods; uses technological devices to support and enhance lectures; uses interesting, relevant, and personal examples; not monotone)
6	*Effective Communicator* (Speaks clearly/loudly; uses precise English; gives clear, compelling examples)
7	*Encourages and Cares for Students* (Provides praise for good student work, helps students who need it, offers bonus points and extra credit, and knows student names)
8	*Enthusiastic About Teaching and About Topic* (Smiles during class, prepares interesting class activities, uses gestures and expressions of emotion to emphasize important points, and arrives on time for class)
9	*Establishes Daily and Academic Term Goals* (Prepares/follows the syllabus and has goals for each class)
10	*Flexible/Open-Minded* (Changes calendar of course events when necessary, will meet at hours outside of office hours, pays attention to students when they state their opinions, accepts criticism from others, and allows students to do make-up work when appropriate)
11	*Good Listener* (Doesn't interrupt students while they are talking, maintains eye contact, and asks questions about points that students are making)
12	*Happy/Positive Attitude/Humorous* (Tells jokes and funny stories, laughs with students)
13	*Humble* (Admits mistakes, never brags, and doesn't take credit for others' successes)
14	*Knowledgeable About Subject Matter* (Easily answers students' questions, does not read straight from the book or notes, and uses clear and understandable examples)
15	*Prepared* (Brings necessary materials to class, is never late for class, and provides outlines of class discussion)
16	*Presents Current Information* (Relates topic to current, real-life situations; uses recent videos, magazines, and newspapers to demonstrate points; talks about current topics; and uses new or recent texts)
17	*Professional* (Dresses nicely [neat and clean shoes, slacks, blouses, dresses, shirts, ties] and no profanity)
18	*Promotes Class Discussion* (Asks controversial or challenging questions during class, gives points for class participation, and involves students in group activities during class)
19	*Promotes Critical Thinking/Intellectually Stimulating* (Asks thoughtful questions during class, uses essay questions on tests and quizzes, assigns homework, and holds group discussions/activities)
20	*Provides Constructive Feedback* (Writes comments on returned work, answers students' questions, and gives advice on test-taking)
21	*Punctuality/Manages Class Time* (Arrives to class on time/early, dismisses class on time, presents relevant materials in class, leaves time for questions, keeps appointments, and returns work in a timely way)
22	*Rapport* (Makes class laugh through jokes and funny stories, initiates and maintains class discussions, knows student names, and interacts with students before and after class)
23	*Realistic Expectations of Students/Fair Testing and Grading* (Covers material to be tested during class, writes relevant test questions, does not overload students with reading, teaches at an appropriate level for the majority of students in the course, and curves grades when appropriate)
24	*Respectful* (Does not humiliate or embarrass students in class, is polite to students [says thank you and please, etc.], does not interrupt students while they are talking, and does not talk down to students)
25	*Sensitive and Persistent* (Makes sure students understand material before moving to new material, holds extra study sessions, repeats information when necessary, and asks questions to check student understanding)
26	*Strives to Be a Better Teacher* (Requests feedback on his/her teaching ability from students, continues learning [attends workshops, etc. on teaching], and uses new teaching methods)
27	*Technologically Competent* (Knows how to use a computer, knows how to use e-mail with students, knows how to use overheads during class, and has a Web page for classes)
28	*Understanding* (Accepts legitimate excuses for missing class or coursework, is available before/after class to answer questions, doesn't lose temper at students, and takes extra time to discuss difficult concepts)

tains 8 items each rated from 1 (*strongly disagree*) to 5 (*strongly agree*). Items addressed each instructor's helpfulness, organization and preparation for the course, ability to motivate students and stimulate their thinking, clarity of teaching objectives and course material, and whether he or she spoke audibly.

Procedure. Four graduate teaching assistants (GTAs) administered the TBC during the last 2 weeks of the Fall 2003 semester. GTAs read a standard set of instructions to their respective sections and provided informed consent information. When students completed the TBC, GTAs collected it and the computer scoring sheets.

The GTAs also distributed and administered the Auburn University standard teaching evaluations within the same week, although not necessarily on the same day. GTAs instructed students to complete the form and return it to them when finished.

Results and Discussion

Initial factor analysis. We combined the TBC data from the four sections and submitted the ratings of the 28 items to an initial factor analysis using the principal components extraction method with no rotation. We conducted this analysis to examine how many factors may be present. We accepted any eigenvalue greater than 1 (Pett, Lackey, & Sullivan, 2003). Three components met this criterion: The first, $\lambda_1 = 13.37$, accounted for 48% of the total variance; the second, $\lambda_2 = 1.58$, accounted for 5.6% of the variance; and the third, $\lambda_3 = 1.20$, accounted for 4.3% of the variance. Monte Carlo studies of factor extractions using random data reveal that expected values for the first three components of a sample of 300 with 30 items are 1.64, 1.56, and 1.48, respectively (Lautenschlager, 1989). Because the first two factors exceeded the values of factors found in random data, it is possible that these two factors reflect the data's true structure. Thus, our initial factor analysis suggested that our data consisted of one, or perhaps two, factors. We conducted a second analysis to examine which solution better explained the data.

Further factor analysis. Our second factor analysis used the maximum likelihood method of extraction with Oblimin nonorthogonal rotation because we expected the two factors to be correlated. We extracted both a one- and two-factor solution, as both seemed plausible based on the previous analysis.

The one-factor solution produced factor loadings for each item ranging from .47 to .79, with most loadings above .6 (see Table 2). This factor encompassed all 28 items. We did not rotate this solution because it involved only one factor. The interpretation of this factor is simply "good teaching" expressed as a unitary construct, with each item tapping some aspect of the construct.

The rotated two-factor solution produced a pattern of loadings in which the first factor corresponded to being "caring and supportive" and the second corresponded to "professional competency and communication skills." We considered single items to load onto a factor if their loading was above .4 (Pett et al., 2003). Using this criterion, 13 items loaded onto the first factor (Items 26, 20, 25, 22, 7, 10, 18, 28, 23, 1, 19, 13, and 8), 11 onto the second factor (Items 4, 6, 14, 3, 12, 27, 15, 2, 24, 21, and 11), and 4 items remained unused (Items 16, 17, 9, and 5). Based on this cutoff point, no items loaded onto both factors (see Table 3). These two factors had a correlation of .73 and thus appear to be closely related, sharing 53% of their variance.

The two-factor solution is similar to Lowman's (1995) model based on interpersonal rapport and intellectual excitement, which appear related to items comprising the caring and supportive and professional competency and communication skills subscales, respectively. According to Lowman, interpersonal rapport is the extent to which teachers interact with their students "in ways that increase their motivation, enjoyment, and independent learning" (p. 27). Intellectual excitement is the clear organization and presentation of course content in a way that emotionally affects students. The two subscales produced by factor analysis of the TBC also bear an interesting link to earlier research using the TBC. Both Buskist et al. (2002) and Schaeffer et al. (2003) found that students and faculty differed along these two dimensions in rating their top 10 qualities and behaviors.

Table 2. Item Loadings in the One-Factor Solution

Item	Total Scale
Humble	.79
Sensitive and persistent	.79
Strives to be a better teacher	.79
Respectful	.76
Encourages and cares for students	.75
Enthusiastic about teaching and about topic	.74
Good listener	.74
Flexible/open-minded	.73
Understanding	.73
Happy/positive attitude/humorous	.73
Approachable/personable	.72
Rapport	.71
Provides constructive feedback	.68
Realistic expectations of students/fair testing and grading	.68
Creative and interesting	.67
Effective communicator	.66
Professional	.66
Knowledgeable about subject matter	.66
Prepared	.63
Punctuality/manages class time	.63
Establishes daily and academic term goals	.62
Promotes critical thinking/intellectually stimulating	.61
Confident	.60
Promotes class discussion	.60
Presents current information	.59
Accessible	.59
Authoritative	.50
Technologically competent	.47

Table 3. Item Loadings in the Two Factor-Solution

Item	Caring and Supportive Subscale	Professional Competency and Communication Skills Subscale
Strives to be a better teacher	.90	
Provides constructive feedback	.84	
Sensitive and persistent	.81	
Rapport	.81	
Encourages and cares for students	.79	
Flexible/open-minded	.65	
Promotes class discussion	.65	
Understanding	.64	
Realistic expectations of students/fair testing and grading	.62	
Accessible	.58	
Promotes critical thinking/intellectually stimulating	.55	
Humble	.55	.30
Enthusiastic about teaching and about topic	.52	.27
Presents current information	.38	.26
Professional	.38	.33
Establishes daily and academic term goals	.37	.30
Confident		.85
Effective communicator		.71
Knowledgeable about subject matter		.71
Authoritative		.63
Happy/positive attitude/humorous	.28	.53
Technologically competent		.48
Prepared	.24	.46
Approachable/personable	.34	.44
Respectful	.39	.43
Punctuality/manages class time	.29	.40
Good listener	.39	.40
Creative and interesting	.34	.39

The one-factor solution is more difficult to interpret as a reflection of general effective teaching. Therefore, we based the remaining analyses on the two-factor solution. We created two new variables by summing the items corresponding to each factor, which produced two subscales: (a) caring and supportive and (b) professional competency and communication skills. We prefer these descriptors to Lowman's (1995) terminology because, based on earlier research using the TBC, our descriptors appear more inclusive of specific factors (behaviors) reflective of these two subscales (cf. Buskist et al., 2002; Schaeffer et al., 2003). We then analyzed these subscales to examine an aspect of construct validity and internal reliability.

Construct validity. To investigate whether these subscales discriminated meaningfully among professors, we performed two one-way ANOVAs across the four professors with the two subscales as dependent measures. The analysis for the first subscale, caring and supportive, produced significant results, $F(3, 307) = 36.59$, $p < .001$. We then performed Tukey post-hoc tests. Professor 1 scored the lowest and was significantly different from the other professors. Professor 2, who scored highest, was different from Professors 1 and 3, but about equal to Professor 4. (We obtained similar results using the more conservative Scheffé and Bonferroni tests.)

The analysis of the second subscale was also significant, $F(3, 308) = 19.11$, $p < .001$. Again, we used Tukey's honestly significant difference to examine the pattern of differences (the interpretation of the results was the same using Scheffé and Bonferroni corrections). Professor 1 scored significantly lower than all other professors, who were approximately equal to each other.

The pattern of results for both subscales was consistent not only with our informal knowledge of these professors' reputations as teachers, but also with students' evaluations of them using the standard Auburn University eight-item teaching evaluation form. We rank ordered the mean of each professor's ratings using this form from the same four classes that also completed the TBC. Professor 1 averaged the lowest overall mean score ($M = 4.30$) and Professors 2, 3, and 4 averaged higher (4.70, 4.73, and 4.69, respectively). Post-hoc comparisons of the students' standard teaching evaluations showed only Professor 1 performing less well than the other professors and did not clearly distinguish the performances of the remaining three. This result is not consistent with our findings on the caring and supportive subscale, which differentiated professors further. However, the students' standard teaching evaluations did match the general pattern of results we obtained using the professional competency and communication skills subscale.

Our examination of the eight items that comprise the standard teaching evaluation explain these data. Six of the eight items reflected course organization and communication skills, whereas only one item directly addressed the instructor's personal style ("Instructor was actively helpful"). The

remaining item could be seen fitting either scale ("Instructor motivated me"). Thus, it is reasonable to assume that the standard teaching evaluation mostly addressed issues similar to those found on the second subscale, hence its similarity to the results of the standard form. Although this evidence is far from conclusive, it does provide initial support for the validity of the two subscales, warranting their further development and investigation.

Reliability. We next examined the internal consistency of the items. The items on the caring and supportive subscale produced an alpha of .93, and items on the professional competency and communication skills subscale produced an alpha of .90. If any item were deleted, these values would decrease. Thus, the items used in both subscales warrant inclusion in their respective subscale. When combined, the alpha for the whole scale equaled .95. Overall, the internal consistency of the scale appears to be excellent.

Study 2

The results of Study 1, although promising, are far from conclusive regarding the TBC's structure. Exploratory factor analysis is an a posteriori technique and, as such, is limited by the data already collected. For this reason, it is not theoretically sound to test hypotheses using exploratory techniques. Therefore, we conducted a confirmatory factor analysis of the TBC designed to compare the one- and two-factor solutions to determine which was a better fit to the data. We also collected data in Study 2 at two different times, midsemester and the end of the semester, to examine the TBC's test–retest reliability.

Method

Participants. We recruited students enrolled across five introductory psychology sections to participate in this study (*n*s = 106, 47, 48, 91, and 31 at Time 1 [313 total]; 90, 65, 57, 79, and 22 at Time 2 [322 total]). A different professor taught each section (four full-time faculty and one part-time instructor). Participants received extra course credit for their participation. Participants ranged in age from 19 to 41 ($M = 19.8$, $SD = 1.7$).

Materials. We modified the evaluative inventory used in Study 1 (consisting of a 5-point Likert-type scale) for use in this study. The first version of this inventory listed *I have no opinion* as the third of five options, which is not a true point on the continuum. We removed this option and moved *sometimes exhibits/has exhibited these behaviors reflective of this quality* to the middle position on the scale. We also moved the first option on the earlier version, *frequently exhibits/has exhibited these behaviors reflective of this quality* into the second position, replacing it with *always exhibits/has exhibited these behaviors reflective of this quality* (see bottom of Table 1). Students used a computer scoring sheet to record their responses and their age.

Procedure. With the assistance of several GTAs, we administered the TBC twice during the Fall 2004 semester: at midsemester and during the last week of the semester. Between the two administrations of the test there was an overlap of 182 students—we used the data provided by these students to determine test–retest reliability. We provided participants with informed consent information and the same set of instructions on the completion of the TBC that we used in Study 1.

Results and Discussion

Confirmatory factor analysis. To evaluate the different models proposed by the exploratory factor analysis in Study 1, we conducted a confirmatory factor analysis of the models using AMOS statistical software. Confirmatory factor analysis is a technique that compares the variability implied in a given model to the variability seen in the data and so may determine how well a given model fits the data. We constructed three models: a one-factor model on which all items loaded, a correlated two-factor model with the items loading in the pattern found in Study 1, and a hybrid model. The hybrid model is a reaction to the idea that correlations reflect underlying factors (the key assumption of factor analysis). Because we allowed both factors in the two-factor model to be correlated, this correlation may represent a higher order factor influencing both. This higher order factor would be conceptually similar to the factor seen in the one-factor model. Hence, the third model is a hybrid between the two other models. Note that we considered the higher order factor in the hybrid model the same "good teaching" factor as the one-factor model; the five items that did not load on either of the two factors loaded on the higher order factor. If the higher order factor is simply added to the two-factor model, it is computationally identical to the correlated two-factor model.

We compared the models at each assessment time separately. We assessed model fit across several indexes of fit: χ^2, the normed fit index (NFI; Bentler & Bonett, 1980), the comparative fit index (CFI; Bentler, 1990), and the root mean square error of approximation (RMSEA; Steiger & Lind, 1980). The standard index of fit is the χ^2 statistic, but it is highly sensitive to sample size and may inflate difference findings (Bentler, 1990). For this reason, confirmatory factor analyses typically include multiple indexes of fit and look for an overall pattern in the results (Thompson, 2004). Each index has differing assumptions regarding the nature of fit, so consensus among multiple indexes is an indication of good fit of a model. Each index has slightly different rules of interpretation. For the χ^2 statistic, values close to the number of degrees of freedom indicate better fit, although the standard *p* values associated with χ^2 are uninterpretable due to such large sample size. For both the NFI and CFI, a value of 1 indicates perfect fit, and any value above .9 is considered acceptable. For the RMSEA, a value of 0 indicates perfect fit, and any value below .1 is acceptable. The values of each index for each model at the two assessment times appear in Table 4.

Table 4. Fit Statistics for the Three Confirmatory Models

	One-Factor Model	Two-Factor Model	Hybrid Model
Midsemester			
χ^2	1340.474	913.213	1169.269
df	350	251	348
NFI	0.963	0.970	0.968
CFI	0.972	0.978	0.977
RMSEA	0.094	0.978	0.977
95% upper bound	0.088	0.084	0.080
95% lower bound	0.099	0.097	0.091
End of the Semester			
χ^2	1158.152	847.129	1043.418
df	350	251	348
NFI	0.970	0.975	0.973
CFI	0.979	0.982	0.982
RMSEA	0.085	0.086	0.079
95% upper bound	0.080	0.080	0.074
95% lower bound	0.091	0.093	0.085

Note. *df* = degrees of freedom; NFI = normed fit index; CFI = comparative fit index; RMSEA = root mean square of approximation.

In examining the fit of the models at midsemester, the models were largely equivalent across the various indexes of fit, although the two-factor model had a lower χ^2 value than the other models. The difference in χ^2 across the models cannot be examined with traditional hypothesis testing because the models are not nested. Therefore, we could not determine that the two-factor model had a significantly lower value of χ^2. One may be tempted to say that the two-factor model was superior to the other two models. Although this point is technically true, the fit of the two-factor model was not markedly superior to the other models, and the difference could be due to chance. Rather, all models fit acceptably well. Although these statistics are a measure of how the model fit overall, it is still possible that parts of the model did not contribute significantly. Fortunately, all paths in all models were statistically significant. Therefore, there were no items or factor relationships that should be dropped from any of the models.

The same pattern of results appeared at the end of the semester. Again, the models were largely equivalent, but the two-factor model had the lowest χ^2 value. All indexes of fit improved at the end of the semester compared to midsemester. The factor analysis that determined the item loadings in the models in Study 1 occurred at the end of the semester as well and so may explain this result. It may be that students use the TBC slightly differently at midsemester or their conception of professors changes to fit with the model as the semester progresses. As at midsemester, all individual path coefficients were statistically significant in all models.

Reliability analyses. We examined the test–retest reliability of the TBC across the two time frames. Given that we generally expected ratings to improve from midsemester to the end of the semester, we expected moderate levels of reliability (i.e., coefficients in the .6 to .8 range). We assessed the reliability of each of the items, the total of the items, and the two scales found in Study 1.

Reliability may be assessed in a variety of ways. The most standard estimation is the Pearson correlation coefficient (r). For the 28 items on the scale, the r values varied from .24 to .64, with most items having coefficients at the .4 or .5 level (19 of the 28 items; p for all item correlations $< .001$). As expected, the reliabilities of the two scales and the total were higher than the value for any individual items. The coefficient for the total scale was .71, $p < .001$. The reliability of the caring and supportive subscale was .68, $p < .001$; the reliability of the professional competency and communication skills subscale was .72, $p < .001$. Overall, the reliability of the scales and total score were at an expected level.

However, the Pearson correlation is an incomplete measure of reliability because it does not take into account the magnitude of change. For example, a person could rate his or her professor with all 1s on the Likert scale at midsemester and then rate the professor with all 5s at the end of the semester, yet still produce a perfect reliability coefficient of 1.0. To account for the magnitude of change, we conducted a regression of the end-of-semester scores using a transformed deviation score of the midsemester scores. Using a deviation score (i.e., the score minus its mean), the intercept of the regression becomes the mean of the dependent variable (i.e., the end-of-semester scores). Thus, the slope of the regression line determines the direction of change from midsemester to end of the semester (i.e., an increase with a positive slope versus decrease with a negative slope) as well as the magnitude of that change (a slope of 1 corresponds to an increase of 1 point on the Likert scale for that item). All slopes were positive for the individual items, with values ranging from .22 to .57 (all ps $< .001$). The total scale had a regression slope of .65, $p < .001$. The caring and supportive subscale had a regression slope of .57, $p < .001$; the professional competency and communication skills scale had a slope of .71, $p < .001$. Overall, evaluations increased about half a point from the midsemester to the end of the semester. The professional competency and communication skills scale seemed to have

a greater increase than the caring and supportive scale. As seen in the correlations, it was also more consistent.

General Discussion

The results of the two studies indicated that the TBC is a psychometrically sound instrument. The first study demonstrated that the TBC had high internal reliability, and the second study indicated that the test–retest reliability was at an expected level, given that scores are expected to increase from midsemester to the end of the semester. The TBC can fit either a one-factor or a two-factor structure. Confirmatory factor analyses indicated that these two models are roughly equivalent, and there is no strong statistical case to select one over the other.

We nevertheless recommend that interested teachers use all 28 items of the TBC and calculate a total score, a score for the first factor, and a score for the second. We did not delete any items from the scale because there was no convincing argument to use the one- or two-factor solution, and so we recommend including both. The deletion of any items will decrease the reliability of the scale, and we found all items to load significantly on the one-factor solution. The total score will give a measure of one's overall teaching; the first factor will be a measure of interpersonal caring and supportive skills, and the second will be a measure of professional competency and communication skills. As of yet, there are no normative or comparison data, so users of the TBC can only compare their performance relative to themselves.

The TBC will likely prove to be of substantial utility to teachers. Our studies did not directly address the potential of the TBC for formative assessment (i.e., to help teachers improve on their weaknesses). Interestingly, though, the TBC will likely be of aid to teachers who wish to improve on a low score on any of the 28 items because items on the TBC have behavioral anchors. Teachers can refer to and adopt these behaviors in an effort to improve their teaching. Other teaching evaluation instruments usually lack such anchors, using instead qualitative personality descriptors that may be difficult for teachers to translate meaningfully to change their classroom behaviors.

The effect of grades on teaching evaluations has been an area of some controversy (e.g., Greenwald, 1997; Marsh & Roche, 1997). It is our hope that the TBC will prove to be more resistant to any effects of student grades on their evaluation of their teacher, largely through the benefit of behavioral anchors on the rating scale. Although we did not evaluate this notion, it may prove to be an interesting area of future study with the TBC.

Our study, although promising, is limited in several ways. First, we studied only teachers of introductory psychology classes. It is reasonable to assume that students may rate teachers in higher level classes differently.

Second, we examined the TBC only in the context of a large research university setting. Students and teachers in different settings such as liberal arts colleges or community colleges may use the TBC differently. However, student and faculty perceptions of master teachers at these sorts of institutions do not differ appreciably (Buskist et al., 2002; Epting et al., 2004; Schaeffer et al., 2003).

Third, we have examined the TBC thus far only within the discipline of psychology, although it may be useful within other disciplines as well. It may well be that not all of the 28 qualities and their behavioral anchors apply equally across disciplines. For example, Item 18 (promotes class discussion) may not be directly applicable to those disciplines that are largely performance based, such as music, art, and theater. Likewise, the sheer size of introductory-level survey courses in some disciplines may prevent teachers from learning students' names, a key behavior reflective of Item 22 (rapport), although there are other ways that rapport might be developed in large classes (Benson, Cohen, & Buskist, 2005; Buskist & Saville, 2004). Although we developed the TBC to apply to classroom instruction at the high school, community college, and college and university level, the utility of all items in assessing teaching may not apply equally across disciplines. How well the use of the TBC will generalize to other disciplines outside of psychology is an empirical question.

Fourth, the TBC consists of qualities and their behavioral counterparts generated wholly by student input. Although some of the behaviors that it recommends as reflective of effective teaching have been found to have solid empirical grounding (e.g., Items 8 [enthusiastic about teaching and about topic] and 19 [promotes critical thinking/intellectually stimulating]; Buskist et al., 2002; Davis, 1993; McKeachie, 2002), other behaviors such as such as "offers bonus points and extra credit" (see Item 7, encourages and cares for students) as yet do not. Again, use of the TBC as an evaluative instrument suggests another topic ripe for empirical study. Clearly, what is known of the TBC is limited at this point, but it is still a psychometrically sound teaching evaluation instrument and thus warrants further study.

References

Baiocco, S. A., & DeWaters, J. N. (1998). *Successful college teaching: Problem-solving strategies of distinguished professors.* Needham Heights, MA: Allyn & Bacon.

Benson, T. A., Cohen, A. L., & Buskist, W. (2005). Rapport: Its relation to student attitudes and behaviors toward teachers and classes. *Teaching of Psychology, 32,* 236–238.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Brookfield, S. D. (1990). *The skillful teacher: On technique, trust, and responsiveness in the classroom.* San Francisco: Jossey-Bass.

Buskist, W., & Saville, B. K. (2004). Rapport-building: Creating positive emotional contexts for enhancing teaching and learning. In

B. Perlman, L. I. McCann, & S. H. McFadden, *Lessons learned: Practical advice for the teaching of psychology* (Vol. 2, pp. 149–155). Washington, DC: American Psychological Society.

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.

Davis, B. G. (1993). *Tools for teaching.* San Francisco: Jossey-Bass.

Epting, L. K., Zinn, T. E., Buskist, C., & Buskist, W. (2004). Student perspectives on the distinction between ideal and typical teachers. *Teaching of Psychology, 31,* 181–183.

Feldman, K. A. (1976). The superior college teacher from the student's view. *Research in Higher Education, 5,* 243–288.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52,* 1182–1186.

Lautenschlager, G. J. (1989). A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. *Multivariate Behavioral Research, 24,* 365–395.

Lowman, J. (1995). *Mastering the techniques of teaching* (2nd ed.). San Francisco: Jossey-Bass.

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,* 1187–1197.

McKeachie, W. J. (2002). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (11th ed.). Boston: Houghton-Mifflin.

Palmer, P. J. (1998). *The courage to teach: Exploring the inner landscape of a teacher's life*. San Francisco: Jossey-Bass.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research.* Thousand Oaks, CA: Sage.

Schaeffer, G., Epting, K., Zinn, T., & Buskist, W. (2003). Student and faculty perceptions of effective teaching: A successful replication. *Teaching of Psychology, 30,* 133–136.

Steiger, J. H., & Lind, J. C. (1980, June). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Wann, P. D. (2001, January). *Faculty and student perceptions of the behaviors of effective college teachers.* Poster presented at the National Institute for the Teaching of Psychology, St. Petersburg Beach, FL.

Notes

1. We thank Alejandro Lazarte and Adrian Thomas for their assistance with statistical analysis and Barney Beins, Drew Christopher, and Steve Davis for reading and commenting on an earlier version of this article.
2. Send correspondence to William Buskist, Psychology Department, Auburn University, Auburn, AL 36849–5214; e-mail: buskiwf@auburn.edu.