Acta Colombiana de Psicología
ISSN: 0123-9155
ISSN: 1909-9711
Universidad Catolica de Colombia

Gerônimo, Vicente Cícero; Oliveira Henklain, Marcelo Henrique; Carmo, João dos Santos; Keeley, Jared Wayne

Utility of the Teacher Behavior Checklist beyond Psychology Majors: Replication with Brazilian Physical Education Students

Acta Colombiana de Psicología, vol. 26, no. 1, 2023, January-June, pp. 214-225
Universidad Catolica de Colombia

DOI: https://doi.org/10.14718/ACP.2023.26.1.14

Available in: https://www.redalyc.org/articulo.oa?id=79874880014

Utility of the Teacher Behavior Checklist beyond Psychology Majors: Replication with Brazilian Physical Education Students

Vicente Cícero Gerônimo Júnior; Marcelo Henrique Oliveira Henklain; João dos Santos Carmo; Jared Wayne Keeley

Cómo citar este artículo:
Gerônimo, V. C. G. J., Henklain, M. H. O., Carmo, J. S., & Keeley, J. W. (2023). Utility of the Teacher Behavior Checklist beyond Psychology Majors: Replication with Brazilian Physical Education Students. *Acta Colombiana de Psicología*, *26*(1), 214-225. https://www.doi.org/10.14718/ACP.2023.26.1.14

Recibido, agosto 29/2021; Concepto de evaluación, junio 14/2022; Aceptado, octubre 04/2022

Vicente Cícero Gerônimo Júnior[1]
ORCID: http://orcid.Org/0000-0002-6713-288x
Federal University of Roraima, Psychology course, Boa Vista, Brazil
Marcelo Henrique Oliveira Henklain
ORCID: https://orcid.org/0000-0001-9884-8592
Federal University of Roraima, Computer Science course, Boa Vista, Brazil
João dos Santos Carmo
ORCID: http://orcid.org/0000-0003-3913-7023
Federal University of São Carlos, Psychology course, São Carlos, Brazil
Jared Wayne Keeley
ORCID: https://orcid.org/0000-0002-1437-5810
Virginia Commonwealth University, Psychology course, Virginia, EUA

Abstract

The Teacher Behavior Checklist (TBC) is a worldwide valued instrument to measure teachers' performance. Nonetheless, the studies about TBC in Brazil are still scarce, with samples mainly composed of psychology and civil engineering students. The aim of this study was to replicate the research by Keeley et al. (2010) to evaluate the psychometric properties of the Brazilian version of the TBC with a new sample. Participants were 107 undergraduates from physical education courses from a Brazilian public university. Participants used the TBC to evaluate three types of teachers: the worst they had ever had, a regular one, and the best one. The order of evaluation of teacher types did not interfere with the response patterns, but as expected, statistically significant differences were found among the three types of teachers. Additionally, the two-factor model of the TBC was confirmed through Confirmatory Factor Analysis, providing additional evidence of construct validity. However evidence to advocate in favor of a one-factor solution was also found. McDonald's Omega results provided evidence of reliability. These findings support the use of TBC in the formative evaluation of teachers in Brazil.
Keywords: Test validity, test reliability, teacher effectiveness evaluation, higher education; Teacher Behavior Checklist.

[1] Psychologist graduated at the Psychology course of the Federal University of Roraima, who produced this research as his completion course work. E-mail: psivicentejr@gmail.com

Utilidad del Teacher Behavior Checklist más allá de la psicología: replicación con estudiantes brasileños de educación física

Resumen

The Teacher Behavior Checklist (TBC) es un instrumento valorado en todo el mundo para medir el desempeño de los profesores. Sin embargo, los estudios sobre el TBC en Brasil siguen siendo escasos, con muestras compuestas principalmente por estudiantes de psicología e ingeniería civil. El objetivo de este estudio fue replicar la investigación de Keeley et al. (2010) para evaluar las propiedades psicométricas de la versión brasileña del Teacher Behavior Checklist (TBC) con una nueva muestra. Participaron 107 estudiantes de educación física de una universidad pública brasileña. Los participantes utilizaron el TBC para evaluar tres tipos de profesores: el peor que hayan tenido, uno regular y el mejor. El orden de evaluación de los tipos de docentes no interfirió con los patrones de respuesta, pero como se esperaba, encontramos diferencias estadísticamente significativas entre los tres tipos de docentes. Además, el modelo de dos factores del TBC se confirmó a través de un análisis factorial confirmatorio, proporcionando evidencia adicional de validez de la construcción. No obstante, también encontramos evidencia para abogar a favor de una solución de un factor. Los resultados del Omega de McDonald indicaron evidencia de confiabilidad. Estos hallazgos apoyan el uso de TBC en la evaluación formativa de los docentes en Brasil.
Palabras clave: Prueba de validez, confiabilidad de la prueba, evaluación de la eficacia docente, enseñanza superior, Teacher Behavior Checklist.

There is a growing concern for evidence-based practice in education (Boysen et al., 2015), which requires evidence-based assessment tools (Andrade & Valentini, 2018) that are particularly scarce in the field of teachers' performance assessment (Henklain et al., 2018). The present study investigates the psychometric evidence of one measure of teaching performance, the Teacher Behavior Checklist (TBC; Buskist et al., 2002).

To the best of our knowledge, the TBC (Buskist et al., 2002) is one of the most prominent instruments worldwide for measuring teachers' performance (Buskist & Keeley, 2018; Henklain et al., 2018). This instrument encompasses 28 teaching qualities and their corresponding behaviors considered typical of excellent teachers (see Buskist et al., 2002 for the complete checklist). Therefore, with the TBC, the degree to which teachers exhibit qualities of excellent teachers is being assessed.

Schneider and Preckel (2017) pointed out that many behaviors of a teacher can promote students' learning and engagement, such as "encouraging and caring for students", "promoting class discussion", "providing feedback", "being friendly", "establishing objectives for learning", etc. Several of these behaviors are covered by the TBC items, which also show many behavioral examples for each quality, helping teachers develop ideas on how to improve their teaching skills. In addition to that, it should be remembered that scientist do not have multiple instruments to assess teachers' performance being studied around the globe. The TBC has this differential, making it possible for educators and researchers to share their knowledge about diverse cultural contexts and educational realities.

Moreover, studies on the TBC have examined its psychometric properties and have contributed to support that it is appropriate for use in teachers' formative assessment. As an example, researchers found concordance in American and Chinese samples of teachers and students regarding their opinion that the TBC qualities are typical of teaching excellence (e.g., Liu et al., 2015). This type of data constitutes evidence of content validity because it indicates that the TBC items adequately represent the spectrum of meanings of the excellent teacher construct. The data also suggest that this validity evidence holds across different cultures.

In this line of psychometric research, two investigations were especially significant for using the TBC as an instrument for measuring teacher performance. The first research, conducted by Keeley et al. (2006) in the USA, investigated the factor structure of the TBC in two studies. In Study 1, an Exploratory Factor Analysis (EFA) found two-factors: "Care and Support" and "Professional competence and communication skills". Then, in Study 2, the factor model, proposed at the end of the first study, was supported by a Confirmatory Factor Analysis (CFA) with a new sample, and a one-factor model. The test-retest reliability of the TBC between the middle and the end of the semester was favorable to the instrument used to evaluate teachers' performance.

In the second investigation, Keeley et al. (2010) assessed the construct validity of the TBC from a new angle. The

researchers employed a technique in which an instrument is used with samples whose results are already known to determine if the instrument correctly measures the construct for which it was designed (for more information, see Cunha et al., 2016). In the case of the TBC, the researchers asked U.S. students to respond to it three times, evaluating their best, worst, and a teacher with whom they studied recently, but who did not stand out as "best" or "worst". Participants were instructed not to imagine an abstract teacher, but a real one with whom they had attended classes. Evidence of construct validity would be favorable only if the scores for each of these types of teachers were different in the expected pattern: best teacher score > recent teacher score > worst teacher score. Participants in two different samples assigned higher scores to the best teachers than to the recent ones, and the recent ones obtained higher scores than the worst teachers. The researchers concluded that students could discriminate between different teachers' performances using the TBC items, which was evidence of construct validity.

In Brazil, studies with the TBC are growing gradually. Henklain et al. (2020) developed an adaptation of this instrument. This version of the TBC retained the 28 items of the original instrument, although some linguistic adjustments were necessary. These researchers also investigated preliminary tests of the instrument's validity and reliability. The participants in this study were predominantly from psychology and civil engineering backgrounds. The results, which they submitted to an EFA, corroborated the two-factor model proposed for the original version of the instrument with some differences related to the items loading on each factor. Reliability tests (Cronbach's alpha of the scale = 0.92, Test-Retest: r_{rho} = 0.75) were also favorable.

In another study, Henklain et al. (in press) conducted a partial replication of the research by Keeley et al. (2010) in which they did not repeat some crucial aspects of the original method, such as requiring each participant to respond to the instrument three times and control the order of responses. Each participant evaluated only one type of teacher they had in college: the best, the worst, or a regular one (neither better nor worst). A CFA corroborated the two-factor model proposed by Keeley et al. (2010), but the authors presented evidence that the TBC could also be interpreted as an unidimensional measure. There was also evidence of construct validity since the best teachers scored higher than the other two types, and regular teachers scored higher than the worst, as expected.

Despite advances in studies with TBC in Brazil, most of these works have used psychology or civil engineering students. Students from other disciplines may view teaching or use TBC differently. Testing TBC with various students is relevant not only for researchers investigating whether the TBC teaching excellence model applies to new contexts and samples, but also to teachers who may teach their disciplines to different majors. This is because they will need some guidance on what teaching excellence might look like beyond psychology or civil engineering related majors.

In this study, it was considered relevant to investigate the use of TBC by physical education students since, as far as it is known, there are no studies with TBC involving this population. In addition to this argument, it was also considered that this major has unique characteristics because the disciplines of education and biology strongly influence it. In contrast, the most researched majors in Brazil are influenced by other subjects that may induce a unique teaching perspective. For example, civil engineering has more physics and mathematics in the curriculum, while psychology, at least in Brazil, is more influenced by philosophy and sociology.

In addition to the sample issue, it should be highlighted that in the Henklain et al. (in press) study, it was not possible to replicate two essential aspects of the Keeley et al. (2010) method, as mentioned above. Since construct validity is the primary type of validity evidence (Cunha et al., 2016), it is crucial to improve the study of Henklain et al. (in press) so that it is possible to increase the empirical basis of support for the Brazilian version of the TBC.

For this reason, the aim of this research was to perform a direct replication of Keeley et al. (2010) work. In this study, the construct validity of the TBC from two complementary angles was investigated: (a) to analyze which model, two-factor (Henklain et al., 2020; Henklain et al., in press) or the one-factor (Henklain et al., in press; Keeley et al., 2006), achieves the best fit with students from different disciplines compared to previous Brazilian TBC studies; and (b) assess whether students' ratings of their worst, regular and best teachers correspond to the expected pattern (best teachers' score > regular teachers' score > worst teachers' score), because if this occurs an important evidence of construct validity would be found. As a secondary objective, an exploratory investigation was initiated, to whether students

evaluations are different when performed with the TBC in the paper and pencil format compared to an online format.

Method

This study is a direct replication (Nosek & Errington, 2017) of Keeley et al. (2010). Therefore, its main methodological aspects were preserved, namely applying the TBC three times to each participant with order control, while testing the TBC with a new sample. The study can be classified as adopting an analytical cross-sectional design (Aggarwal & Ranganathan, 2019) in which students had to rate three types of teachers – best, worst, and regular– using the paper and pencil version of the TBC adapted to Brazilian Portuguese (Henklain et al., 2020). These data were analyzed to investigate TBC's psychometric properties. Additionally, at the end of the study, participants were asked to rate the same teachers again, but now using an online version of the TBC. The objective was to test the degree of correlation between the TBC scores in the paper and pencil version and in the online version to gather preliminary evidence about the possibility of using both versions of the instrument in Brazil.

Participants

The study used a sample by convenience, and all the students had to sign a consent form to be enrolled in it. Participants were 107 Physical Education students from a public university in the Brazilian state of Roraima (corresponding to approximately 43% of the number of undergraduate students enrolled in this university course), with 57 women and 50 men, and a mean age of 24.9 years ($SD = 5.5$). To participate in this study, it was requested for students to have sufficient experience with university teachers. For this reason, only students from the second module or higher could participate. One module is equivalent to one semester with three and a half months of class time. The full degree in Physical Education consists of eight modules.

The sample was obtained from the following semesters: 20 students from the second semester, 10 from the third semester, 12 from the fourth semester, 20 from the fifth semester, 12 from the sixth semester, 14 from the seventh semester, and 19 from the eighth semester. Only three participants reported having a disability; 56.1% identified themselves as belonging to the middle social class, 43.9% as low-income, and no participant reported belonging to the high-income social class. Thirty-three students participated in a two-week reapplication of the TBC, now using an online version of it (30.84% of the initial sample).

Instruments

The research protocol consisted of two instruments: (1) three copies of the Brazilian version of the Teacher Behavior Checklist (TBC) (in paper and pencil format), and (2) a copy of a demographic questionnaire that assessed students' gender, age, disability, social class, and course module. The TBC was adapted by Henklain et al. (2020) through a translation procedure by independent translators, followed by investigation of semantic and content validity, and finalized with back-translation. This instrument has 28 items (teaching qualities and corresponding behaviors) rated on a five-point frequency scale, "1 = *never exhibits*" to "5 = *always exhibits*". Sample item: "Accessible/available (informs of work schedule; facilitates schedule to see students; makes available telephone number, WhatsApp, and e-mail contact; responds to student contact)".

The Cronbach's alpha of the TBC Brazilian version found by Henklain et al. (2020) was .92 (omega = .94), and an Exploratory Factor Analysis (EFA) showed that the TBC could be interpreted by a two-factor model: Relational Behaviors (Factor 1, alpha = .85, omega = .89) and Pedagogical Behaviors (Factor 2, alpha = .90, omega = .92). Henklain et al. (in press) confirmed the two-factor model by Confirmatory Factor Analysis (CFA, TBC's alpha = .96, Factors' 1 and 2 alphas = .93, TBC's omega = .97, Factor's 1 omega = .95, Factor's 2 omega = .91), but advocated in favor of a one-factor structure as probably the most appropriate and capable for measuring effective teaching considering the closeness to unidimensionality assessment. Henklain et al. (2020) also found evidence of temporal stability based on a test-retest procedure ($r_s = .748$, *p(one-tailed)* $< .01$, $N = 229$; Factor 1: $r_s = .59$, *p(one-tailed)* $< .01$, $N = 229$; Factor 2: $r_s = .75$, *p(one-tailed)* $< .01$, $N = 229$).

An online version of the TBC was created, containing the same items as the paper and pencil format, which could be accessed by a Google Forms link. It was found that the TBC used to evaluate the worst teachers showed an alpha of .95, the one used to evaluate regular teachers showed

an alpha of .95, and the TBC related to the best teachers exhibited an alpha of .97.

Procedure

The educational institution, teachers, and university students were informed of the objectives of this study. Data were collected inside a classroom without the presence of a teacher. Participants were given the following instructions: "You will answer the TBC three times, each time thinking about a different teacher you had in college. Here is an example of what might happen: The first time, you will think about the worst teacher you had in college, and then you will evaluate each item on the TBC thinking exclusively about this teacher. Next, you will think about the best teacher you ever had and evaluate each TBC just considering this teacher. Finally, you should choose a regular teacher and evaluate him or her. It can be any teacher if he/she is neither the best nor the worst. Each person will do these three evaluations but in different orders".

To avoid student confusion, the TBC protocols for evaluating each type of teacher were printed in different colors: blue for the best teachers, white for the regular ones, and red for the worst. In addition to the colors, all protocols had specific instructions on the type of teacher to be evaluated. Each protocol had a code that the participant had to register for use in the reapplication of the TBC in the online version. For the reapplication, the TBC online version was implemented in Google Forms. The order of rating the best, worst, and the regular teacher was randomized. To ensure the largest possible sample, data collection was also conducted individually for students who were not in class at the time of the first data collection.

At the end of the completion of the three TBC ratings, the participants answered the demographic questionnaire and, finally, the researcher thanked them and provided the following instruction: "Within 15 days, I am going to send you a Google Forms link so that you can evaluate again the same teachers you evaluated today, in the same order". The second contact with the students was made exclusively via WhatsApp to remind them of the last phase of the research. During this period, the students were participating in an internship, so they no longer had face-to-face meetings at the university.

Compliance with Ethical Standards

Conflict of interest

The authors declare they have no conflict of interest.

Funding

There was no specific funding assigned to this project, but the present study is part of the research program of the National Institute of Science and Technology about Behavior, Cognition and Teaching (INCT-ECCE).

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The research was approved by the Brazilian platform for ethical committees (Brazilian Platform, CAAE 54448416.6.0000.5302).

Informed consent

Informed consent was obtained from all individual participants included in the study.

Data analysis

Statistics were calculated with the R software (Development Core Team, 2017). The mean score and standard deviation in the total TBC score were calculated for each participant in relation to the three types of teachers. Next, the Shapiro-Wilk normality test was applied to the raw TBC data for each item and type of teacher. It was found that, individually, the items of the three types of evaluation (best, worst, and regular) were not normally distributed (all with $p < .05$). Considering that an asymmetric distribution for each TBC item was found, that the sample was defined by convenience, and that the scale was ordinal, it was decided to use non-parametric statistics to perform hyphotesis test and correlation calculations.

A confirmatory Factor Analysis (CFA) was performed based on the two-factor model proposed by Henklain et al. (2020) and the one-factor model suggested by Henklain et al. (in press). The adopted fit indexes and criteria are indicated in Table 2, based on Hair et al. (2009). The assessment of the closeness to unidimensionality assessment (Damásio & Dutra, 2017) was also performed using Factor software (version 10.5.03, Ferrando

& Lorenzo-Seva, 2017). Based on Volpato and Barreto (2011) and Field (2009), Friedman's non-parametric test was used to evaluate whether the mean TBC score was different for each type of teacher, which accounts for the dependence between the measures of the study, as the same participant used the TBC three times in a sequence. The same test was performed to assess whether the order of each type of TBC influenced the way participants responded to the instrument. If a significant *p*-value was found, Nemenyi's multiple comparison *posthoc* tests were used, adopting a Bonferroni correction. Finally, Cronbach alpha and McDonald's Omega of the scale and a Spearman correlation were calculated to test the association between TBC scores in the paper-and-pencil format and TBC in the Google Forms format. According to Field (2009), there is better evidence of a correlation between two variables when its magnitude exceeds |.3| and is statistically significant.

Results

Considering the data collected with the TBC's paper-and-pencil version, the CFA results from the one-factor and two-factor solutions were compared. Table 1 shows the fit indexes for both models.

As noted in Table 1, both the one-factor and the two-factor models based on the present study's data provided good fit indexes. The two-factor model showed slightly better fit indexes when compared to the one-factor solution. An adequate χ^2/df ratio was found with the sample for both factorial solutions, which was not obtained by Henklain et al. (in press). To better investigate which model is the best, the assessment of closeness to unidimensionality was performed by adopting the cut-off points suggested by Damásio and Dutra (2017).

The value of UniCo (Unidimensional Congruence) was 0.99, which is higher than 0.95 cut ppoint. This result suggests that the data can be treated as essentially unidimensional. The ECV (Explained Common Variance) value was 0.926, which is greater than 0.85 and, again, suggests that TBC is a unidimensional measure. The MIREAL (Mean Item REsidual Absolute Loadings) value was 0.189, being less than 0.4, which is also considered evidence that the data can be treated as essentially unidimensional. The I-Unico (Item Unidimensional Congruence) was examined too, and it was found that only Item 3 was lower than 0.95. When considering the I-ECV (Item Explained Common Variance), six items lower than 0.85 were obtained: Item 3 (0.647), Item 12 (0.842), Item 15 (0.813), Item 17 (0.833), Item 21 (0.836), and Item 28 (0.796). When analyzing the I-REAL (Item REsidual Absolute Loadings), there was only one item lower than 0.4 (Item 3, 0.431). Therefore, for most items the indicators suggest that the one factor solution is the most appropriate to interpret the TBC results.

The one-factor solution fit indexes were adequate, and this solution is theoretically reasonable. It has been recommended by Keeley et al. (2006) and Henklain et al. (in press). As the evidence from the closeness to unidimensionality assessment suggests that the TBC is essentially unidimensional, it was decided to use the one-factor solution to analyze this paper´s data. Nonetheless, Tables 2 and 3 show the *lambda* values, standard errors, *z*-scores, and *p*-values for the 28 TBC items organized into one and two-factors solutions.

Table 1
Comparison of the CFA results for one and two-factor models

Indexes	Cut points	One-factor	Two-factor
χ^2	---	850.236, $p < .001$	854.234, $p < .001$
dF	---	275	349
χ^2/dF	> 2 and < 5	3.092	2.45
GFI	≥ 0.90	0.994	0.996
AGFI	≥ 0.90	0.992	0.994
SRMR	< |0.08|	0.056	0.050
RMSEA	< 0.1	0.081 (CI 90%: 0.075-0.087)	0.067 (CI 90%: 0.062-0.073)
CFI	≥ 0.95	0.996	0.997
TLI	≥ 0.95	0.996	0.997

Note. χ^2 = chi-square; dF = degrees of freedom; *p* = *p-value*.

Table 2

Results of the CFA performed for the one-factor model

Items	λ	SE	*z*	*p*
1	0.764	0.026	29.513	*
2	0.838	0.019	44.776	*
3	0.641	0.034	19.047	*
4	0.868	0.016	54.233	*
5	0.883	0.015	59.402	*
6	0.874	0.016	53.929	*
7	0.877	0.016	56.573	*
8	0.871	0.015	57.035	*
9	0.890	0.013	66.261	*
10	0.809	0.021	38.747	*
11	0.874	0.015	58.581	*
12	0.773	0.024	31.660	*
13	0.813	0.022	37.748	*
14	0.826	0.021	40.226	*
15	0.853	0.017	49.337	*
16	0.845	0.018	48.274	*
17	0.754	0.026	28.866	*
18	0.859	0.017	49.556	*
19	0.881	0.014	61.635	*
20	0.868	0.016	54.490	*
21	0.788	0.024	33.503	*
22	0.819	0.020	40.976	*
23	0.818	0.020	40.936	*
24	0.798	0.022	35.688	*
25	0.749	0.025	29.874	*
26	0.764	0.026	29.513	*
27	0.838	0.019	44.776	*
28	0.641	0.034	19.047	*

Note. λ = Lambda; SE = Standard-error; z = z-score; *p* = p-value; * = $p < .001$.

All *lambdas* were statistically different from zero ($\lambda \neq 0$, $z > 1.96$, $p < .001$), varying between 0.641 (Item 3) to 0.89 (Item 9) in the one-factor solution, and 0.642 (Item 3) to 0.89 (Item 7 and Item 25) in the two-factor solution. The fact that each item loaded significantly in the expected direction is additional evidence that both models are adequate for the TBC data in Brazil, even though it is believed that the TBC is essentially unidimensional.

Considering the scores calculated for the one-factor solution, the existence of order effects between the presentation of the best, worst, and regular teacher ratings was tested; there were none[1] ($\chi^2(5) = 9.832$, *ns*, $W = 0.772$). The data about the order effect was plotted in a boxplot where it was also not found striking differences among the six orders. There was, in turn, a statistically significant difference between the three types of teachers ($\chi^2(2) = 186.19$, $p < .001$, $W = 0.887$). To identify where the differences were, a Nemenyi multiple comparison test was performed showing that all comparisons were significant (*p*s <.001),

[1] Summary statistics for order effects are available from the authors upon request.

Table 3
Results of the CFA performed for the two-factor model

Factor	TBC Items	λ	SE	*z*	*p*
Factor 1 (Relational Behaviors)	1	0.779	0.026	30.212	*
	2	0.853	0.018	47.426	*
	7	0.899	0.015	60.511	*
	10	0.828	0.020	41.080	*
	11	0.895	0.014	62.886	*
	12	0.782	0.024	32.375	*
	13	0.836	0.020	41.337	*
	17	0.775	0.026	30.051	*
	22	0.832	0.019	42.737	*
	23	0.838	0.020	42.336	*
	24	0.815	0.022	37.525	*
	28	0.766	0.024	31.329	*
Factor 2 (Pedagogical Behaviors)	3	0.642	0.034	18.790	*
	4	0.870	0.016	53.730	*
	5	0.888	0.015	60.468	*
	6	0.874	0.016	53.246	*
	8	0.875	0.015	57.401	*
	9	0.893	0.013	67.152	*
	14	0.828	0.021	39.910	*
	15	0.858	0.017	50.588	*
	16	0.855	0.017	50.746	*
	18	0.862	0.017	49.733	*
	19	0.882	0.014	61.158	*
	20	0.877	0.015	57.283	*
	21	0.794	0.023	34.378	*
	25	0.899	0.013	69.112	*
	26	0.875	0.016	55.348	*
	27	0.753	0.026	29.440	*

Note. λ = Lambda; SE = Standard-error; *z* = z-score; *p* = p-value; * = $p < .001$.

worst ($M = 2.4$, $SD = 0.6$, $Mdn = 2,39$) *versus* best ($M = 4.6$, $SD = 0.3$, $Mdn = 4,64$), worst *versus* regular ($M = 3.6$, $SD = 0.7$, $Mdn = 3,54$), and best *versus* regular. As expected, the scores of the best teachers were higher than the other two, and the regular teachers obtained a higher score than the worst. The same pattern was found for best, regular, and worst teachers, when these tests reported here were performed specifically with the data of each one of the six orders. This suggests that the variable type of teacher influences the score on TBC and that the order of assessment does not.

The three Cronbach alphas calculated for the whole scale and with data organized by each type of teacher evaluated, showed excellent results: whole scale = .98; worst teacher α = .91; regular teacher α = .94; best teacher α = .90. The same pattern was found with the McDonald's Omega: whole scale = .98; worst teacher α = .93, regular teacher α = .95; best teacher α = .92.

Finally, a weak and statistically significant (or marginal in the case of Factor 2) correlation ($rho = .20$, $p = .041$, $N = 33$; Factor's 1 rho = .20, $p = .043$; Factor's 2 rho = .20,

p = .050) was found between the two TBC formats. In this final phase of the study, a very low adherence of the participants was experienced, since only 33 returned to answer the TBC considering each type of teacher (best, worst, and regular).

Discussion

The main objective was to conduct a direct replication of Keeley et al. (2010) study to investigate the TBC construct validity from two complementary angles:

i. To analyze the data with a CFA to verify which factorial solution, one or two factors, would be the most adequate for students from a different academic discipline, based on previous studies with the TBC in Brazil (Henklain et al., 2020; Henklain et al., in press).
ii. To investigate if the scores obtained in students' evaluations with different types of teachers (best, regular, and worst) would follow the expected pattern of results.

Finally, as a secondary objective, an investigation was carried out to verify whether using the TBC in a different format would change the initial assessments. It was found that the present study successfully replicated Keeley et al. (2010), providing additional evidence of construct validity for the TBC. First, the study provided evidence of construct validity by means of good fit indexes for the one- and two-factor models, considering data from physical education students that have not been studied before in Brazil. In analyzing the factorial solutions, it was decided to use the one-factor solution mainly because the closeness to unidimensionality assessment (Damásio & Dutra, 2017) pointed out that the TBC is essentially a unidimensional measure.

In addition to that, some advantages of the one-factor model should be considered: (a) it makes sense to theoretically surpass the two-factor model, considering that the division of teachers' qualities into two distinct categories is more didactic (e.g., useful for teaching about what constitutes excellent teaching, and for giving feedback to teachers) than tangible; (b) it is easier to analyze and interpret data; (c) it makes easier to compare TBC data across countries to investigate universal principles of teaching and formative assessment (Buskist & Keeley, 2018).

In fact, the one-factor solution was previously recommended by Keeley et al. (2006) and Henklain et al. (in press) as a possible approach to analyze data collected with the TBC. These two studies (Henklain et al., in press; Keeley et al., 2006) did not find strong statistical reasons to recommend one factor solution over the other and suggested that evaluation should focus on what is most appropriate considering the objectives thereof. For instance, perhaps the two-factor solution could give us a better understanding of teachers' performance, and more information to prepare feedback, as the study could specifically address their relational and pedagogical behaviors (as proposed by Henklain et al., 2020). Nonetheless, with the one-factor model, a robust score of effective teaching can be obtained to analyze, from a broad perspective, the performance of several teachers from one or more educational institutions.

When comparing teachers' performance, our data show that the scores of the worst, best, and regular teachers are statistically different in the expected pattern. It is important to remember that construct validity is the main psychometric property or, at least, one of the most basic pieces of evidence one must find to show that an instrument is adequate for use because it measures what was designed to (Cunha et al., 2016). It is important to emphasize that few TBC studies investigate its construct validity, most of them are mainly concerned with content validity and descriptions of what teachers and students think about excellence in teaching (Buskist et al., 2002; Buskist & Keeley, 2018; Liu et al., 2015). Although these studies are important, further research should be conducted to study the psychometric properties of crucial relevance for the use of TBC as an instrument in natural contexts, such as the classroom.

The results also showed that the favorable psychometric evidence found were not influenced by the order in which the three types of TBC were presented, which also confirmed the study's initial expectations. The fact that prior TBC ratings did not influence students' subsequent scores on the instrument generates confidence that it could be used for evaluative purposes in applied settings. This finding suggests that TBC ratings are specific to the teacher being rated and not skewed by students' recollections of other teachers. The Cronbach's alpha and the McDonalds' Omega calculated for the three types of TBC (worst, best, and regular) and the whole scale were excellent, suggesting good reliability (according to Field, 2009).

Finally, the correlational investigation conducted with 30.84% (N = 33) of the initial sample showed a weak association between the two versions of the TBC. Perhaps this result occurred because the participants were exhausted from

having to answer the TBC again and were not so attentive or careful while performing the task. It should be considered that answering the TBC three times, as was requested in the research, means filling in 84 items, which can be a burden to most people. Another hypothesis for the weak correlation is that there are some psychometric properties that are different among the two TBC formats tested.

To confirm any of these hypotheses, further research would have to be conducted on this question. One caveat about the present result is that the electronic format of TBC was tested with less than a half of our original sample, which occurred because the participants had no interest in remaining engaged in the research even though they invited. Nonetheless, a positive and statistically significant correlation was found suggesting a promising path for future investigations. It is likely that the TBC could be used in both formats, paper-and-pencil and online. Theoretically, there is no reason to believe that the students' interpretation and use of the TBC items would be any different in these two formats, and the present correlation is the first data that raises that question. The online format is easier to administer and analyze, which is very important considering all the social and educational challenges imposed by the COVID-19 pandemic. Therefore, researchers interested in TBC have a fruitful line of research to work on. This path encompasses studies about measurement invariance between different TBC formats and groups (Damásio, 2013), which are necessary to advance the comparison of TBC data between different cultures, groups or social contexts.

Altogether, this study broadened the examination of the TBC in Brazil to a new institution and discipline, which is helpful because the main primary data thus far have come from psychology and civil engineering courses from only one institution (Henklain et al., 2020; Henklain et al., in press). It is encouraging to find the same results with ratings of Physical Education students. The present findings suggest that the TBC may be adequate to evaluate teachers from different disciplines, making it very useful to teachers, educational institutions, researchers, and policy makers that are trying to understand what works best in education. As Hattie (2015) pointed out, it is very important for educational success that teachers have a common view of what effective teaching looks like. Therefore, as a valued worldwide measure of teachers' performance, the TBC has this potential to assist in describing what should be prioritized to guarantee effective teaching.

Beyond its contributions, this study had some limitations that should be considered. Firstly, the sample was small, compared to the sample sizes typically included in psychometric works. This problem limits the generalization to other Physical Education students. However, the sample had adequate power to detect significant effects of the study, since it represented a substantial portion (43%) of the students in the course who could have been recruited. In addition to this fact, it should be noted that the small sample is related to the difficulty of convincing students to participate in a research in which they are asked to respond to an instrument three times, having to fill in many items. A second caveat is that the sample used was defined by convenience, which also limits the generalization of the results to Brazilian undergraduate Physical Education students. However, the fact that the results were similar to students from other institutions and academic disciplines indicates that the current results should be trusted.

Future studies should continue to investigate the TBC psychometric properties in Brazil and other countries, trying to select representative samples of the undergraduate population, and expanding the courses and institutions to test the TBC psychometric properties under various conditions. In conclusion, this study found that the Brazilian version of the TBC performed well with a new group of students and institution. It measured the quality of teaching in an expected pattern, providing evidence of construct validity. It was stable over time, and the factor structure found in other studies was replicated in this new context. The one-factor solution was also found to be particularly adequate for analyzing TBC data and should be adopted. These conclusions lead to believe that the TBC appears to be a useful and adequate measure for assessing teaching quality in Brazil.

References

Aggarwal, R., & Ranganathan, P. (2019). Study designs: Part 2 – Descriptive studies. *Perspectives in Clinical Research*, *10*(1), 34-36. https://doi.org/10.4103/picr.PICR_154_18

Andrade, J. M. de, & Valentini, F. (2018). Diretrizes para a construção de testes psicológicos: a Resolução CFP n. 009/2018 em destaque [Guidelines for the Construction of Psychological Tests: Regulation CFP No: 009/2018 in

Highlight]. *Psicologia: Ciência e Profissão*, *38*(especial), 28-39. https://doi.org/10.1590/1982-3703000208890

Boysen, G. A., Richmond, A. S., & Gurung, R. A. R. (2015). Model teaching criteria for psychology: initial documentation of teacher's self-reported competency. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 48-59. https://doi.org/10.1037/stl0000023

Buskist, W., & Keeley, J. W. (2018). Searching for universal principles of excellence in college and university teaching. *New Directions for Teaching and Learning*, *156*, 95-105. https://doi.org/10.1002/tl.20321

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. 30-39). New York: Psychology Press.

Cunha, C. M., Neto, O. P. de A., & Stackfleth, R. (2016). Principais métodos de avaliação psicométrica da validade de instrumentos de medida [Main methods of psychometric evaluation of the validity of measuring instruments]. *Revista de Atenção a Saúde*, *14*(47), 75-83. https://doi.org/10.13037/ras.vol14n47.3391

Damásio, B. F. (2013). Contribuições da Análise Fatorial Confirmatória Multigrupo (AFCMG) na avaliação de invariância de instrumentos psicométricos [Contributions of the Multigroup Confirmatory Factor Analysis in the invariance evaluation of psychometric tests]. *Psico-USF*, *18*(2), 211-220. https://doi.org/10.1590/S1413-82712013000200005

Damásio, F. B., & Dutra, D. F. (2017). Análise fatorial exploratória: Um tutorial com o software Factor [Exploratory factor analysis: A tutorial with the Factor software]. In B. F. Damásio, & J. C. Borsa (orgs.), *Manual de desenvolvimento de instrumentos psicológicos* (pp. 241-265). Vetor.

Development Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development, and future directions. *Psicothema*, *29*(2), 236-241. http://dx.doi.org/10.7334/psicothema2016.304

Field, A. (2009). *Descobrindo a estatística usando o SPSS* [Discovering Statistics Using SPSS] (2nd ed.). Artmed.

Hair, J. F. Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados* [Multivariate Data Analysis] (6th ed.). Bookman.

Hattie, J. (2015). The applicability of visible learning to higher education. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 79-91. http://dx.doi.org/10.1037/stl0000021

Henklain, M. H. O., Carmo, J. S., Haydu, V. B., Muniz, M., Buskist, W., & Keeley, J. W. (2020). Teacher Behavior Checklist: Psychometric evidence in teacher evaluation by Brazilian university students. *Paidéia*, *30*(e3025), 1-11. https://doi.org/10.1590/1982-4327e3025

Henklain, M. H. O., Muniz, M., Carmo, J. S., Haydu, V. B., Keeley, J. W., & Buskist, W. (in press). Teacher Behavior Checklist's psychometric properties: A study with Brazilian undergraduates. *CES Psicología*.

Henklain, M. H. O., Carmo, J. S., & Haydu, V. B. (2018). Contribuições analítico-comportamentais para descrever o repertório de professores universitários eficazes [Behavior-analytical contributions to describe the repertoire of effective university teachers]. *Revista Brasileira de Orientação Profissional*, *19*(2), 197-207. http://dx.doi.org/1026707/1984-7270/2019v19n2p197

Keeley, J. W., Furr, R. M., & Buskist, W. (2010). Differentiating psychology students' perceptions of teachers using the teacher behavior checklist. *Teaching of Psychology*, *37*, 16–20. https://doi.org/10.1080/00986280903426282

Keeley, J. W., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology*, *33*(2), 84-91. https://doi.org/10.1207/s15328023top3302_1

Liu, S., Keeley, J., & Buskist, W. (2015). Chinese College Students' Perceptions of Excellent Teachers Across Three Disciplines: Psychology, Chemical Engineering, and Education. *Teaching of Psychology*, *43*(1), 83-86. https://doi.org/10.1177/0098628315620888

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in Cancer Biology: Making sense of replications. *eLife*, *6*(e23383), Article e23383. https://doi.org/10.7554/eLife.23383

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565-600. https://doi.org/10.1037/bul0000098

Volpato, G. L., & Barreto, R. E. (2011). *Estatística sem dor!!!* [Statistics without pain!!!]. Botucatu: Best Writing.