Development and Validation of the Online Teaching Effectiveness Scale

Elizabeth Reyes-Fournier, Edward J. Cumella, and Gabrielle Blackman
Purdue University Global

Michelle March
College of Lake County

Jennifer Pedersen
University of Alaska Anchorage/Kenai Peninsula College

Abstract
The currently available measures of online teaching effectiveness (OTE) have several flaws, including a lack of psychometric rigor, high costs, and reliance on the construct of traditional on-the-ground teaching effectiveness as opposed to the unique features of OTE (Blackman, Pedersen, March, Reyes-Fournier, & Cumella, 2019). Therefore, the present research sought to establish a psychometrically sound framework for OTE and develop and validate a measure based on this clearly-defined construct. The authors developed pilot questions for the new measure based on a comprehensive review of the OTE literature and their many years of experience as online instructors. Students enrolled in exclusively online coursework and programs at Purdue University Global (N = 213) completed the survey, rating the effectiveness of their instructors. Exploratory Factor Analysis produced four clear OTE factors: Presence, Expertise, Engagement, and Facilitation. The resulting measure demonstrated good internal consistency and high correlations with an established OTE measure; good test-retest reliability; and predictive validity in relation to student achievement. Confirmatory Factor Analysis revealed a good fit of the data and yielded a final 12-item OTE measure. Further refinement and validation of the measure are recommended, particularly with students in other universities, and future research options are discussed.

Keywords: online teaching effectiveness, instructor effectiveness, distance learning, student evaluations, asynchronous learning

Reyes-Fournier, E. Cumella, E.J., March, M., Pederson, J. & Blackman, G. (2020). Development and validation of the online teaching effectiveness scale. *Online Learning, 24*(2), 111-127. https://doi.org/10.24059/olj.v24i2.2071

Development and Validation of the Online Teaching Effectiveness Scale

By the fall of 2015, there were nearly six million students enrolled in online courses at colleges and universities in the United States. Half were exclusively online students (National Center for Education Statistics, 2018). The advent of online education, in conjunction with Internet access, has ushered in technical advances in instructional design and teaching platform innovations (Nilson & Goodson, 2017). Despite this growth, the assessment of online courses and instructors has not kept up with the developments (Berk, 2013). Many tools now used to assess online instructors' teaching skills were developed for on-the-ground teaching (Thomas & Graham, 2017). In essence, in terms of assessment, online teaching effectiveness has been treated as a virtual extension of on-the-ground teaching effectiveness, rather than as the unique instructional phenomenon that it is (Blackman, Pedersen, March, Reyes-Fournier, & Cumella, 2019).

The Seven Principles of Good Practice in Undergraduate Education (Chickering & Gamson, 1987), a list of best practices for postsecondary teaching, has been a gold standard in conceptualizing higher education for over 30 years. These best practices were derived through a review of the literature rather than an established factorial processes (Chickering & Gamson, 1987). Thomas and Graham (2017) noted that most current online teaching effectiveness (OTE) measures use these seven principles for evaluating online instructors with little attempt at adaptation for or validation within online milieus. The lack of research and appropriate tools for online teaching evaluation remain pressing concerns for online instructors and administrators. In the Instructional Technology Council's 2016 *Annual National eLearning Report*, online university staff and administrators reported that adequate assessment of eLearning classes was one of their greatest challenges. Annual survey results have indicated for four years running that evaluation of online faculty remains a top concern for online university administrators (Lokken, 2016). Thus, proper measurement of OTE needs to be addressed using appropriate qualitative and quantitative research strategies (Lokken, 2016; Serdyukov, 2015).

For instance, commercially available measures, such as the Electronic Student Instructional Report II (e-SIR II) offered by Educational Testing Services (Klieger et al., 2014; Pike, 2004), are available for evaluating distance education. Nonetheless, the e-SIR II was adapted from the SIR II, which is used specifically for on-the-ground teacher evaluations (Centra, 2006). The e-SIR II retains more than half the items from the SIR II and then includes new items created by a panel of online educators. The e-SIR II specifically addresses the following dimensions: five items pertaining to planning and course organization; five to the interaction between faculty and students; five to specific course activities, such as grading, exams, and assignments;, eight to the teacher's instruction and course material; five to course outcomes; three to student effort and engagement; and three to the amount of work, pace, and difficulty of the course (Liu, 2011). Of these 40 items, more than half are specific to course content, such as materials, subject matter, assignments, and exams, which are often not controlled by the online instructors and therefore not relevant to rating OTE. Efforts to validate the e-SIR II demonstrate that, of the measured domains, the highest correlations between factors were student effort and course outcome, r =.78 (Liu, 2011). Thus, as it pertains to OTE, these issues raise questions about whether students are evaluating the course, their own effort, or the efficacy of their instructors.

Of the other available OTE measures, Bangert's 26-item Student Evaluation of Online Teaching Effectiveness (SEOTE; 2006, 2008) is purported to focus on producing valid feedback to instructors in regard to the effectiveness of their teaching. The SEOTE was initially founded in Chickering and Gamson's (1987) principles of effective teaching, but after initial exploratory

factor analysis, the measure produced a four-factor model of OTE (Bangert, 2008). The four factors measured by the SEOTE include "student-faculty interaction, active learning, time on task, and cooperation among students" (Bangert, 2008, p. 25). Of these factors, only one speaks directly to the work of the online instructor. Of the 26 items on the SEOTE, 14 are written with the following introductory phrase, "This course…", referencing course design and not directly measuring the effectiveness of the instructor (Bangert, 2008). Because online instructors often have little input into course design, items of this kind do not adequately reflect OTE.

Blackman et al. (2019) have reviewed available OTE measures and concluded that each has significant limitations. In summary, attempts to create a measure for OTE have tended to focus on evaluating the course and student effort rather than the effectiveness of the instructor. Thus, the existing measures use inappropriate constructs and questions for the online mode of teaching. All measures lack sufficient demonstrated psychometric rigor, including a lack of reliability and validity data. In the absence of appropriate tools to measure OTE, online universities have in many cases bypassed the role of instructor effectiveness and focused their quality improvement efforts on their technological delivery systems and students' digital experiences (Serdyukov, 2015). The lack of assessment of OTE clearly compromises the ability of online educators and administrators to fully assess their educational programming and engage in quality improvement initiatives. Therefore, an evidence-based, publicly available measure of OTE, based on current OTE research, would appear to contribute to both research and practice.

In short, the construct of OTE has not been well-defined or established through research. Rather, the face-to-face teaching paradigm has been imported with little modification and applied to OTE, even though online teaching diverges from face-to-face teaching in substantive ways. Therefore, we ask:

1. What factors empirically comprise the construct of OTE?
2. Does an empirically-derived OTE measure demonstrate adequate reliability and validity?

Method

Participants

Participants were recruited from undergraduate and graduate classes at Purdue University Global (PG), representing all undergraduate grades and all levels within master's programs. Classes represented seven academic departments at PG's College of Social and Behavioral Sciences: Communication, Criminal Justice, Early Childhood Education, Fire Science, Human Services, Psychology, and Social Science. The study was approved by the Dean of the College, each academic department chair, and the PG Institutional Review Board.

Department chairs emailed their faculty about the research opportunity. The email assured faculty that their decisions to participate or not would never be known by their chair or anyone besides the researchers; that their decisions to participate were voluntary and would have no impact, either adverse or positive, on their work at the university. Inclusion criteria for students consisted of being 18 years-old or older and enrolled as a student at recruitment time. Exclusion criteria consisted of being less than 18 years-old and not being enrolled. During week 8 of their courses, all students in the selected classes received emails containing a research announcement and video link. Week 8 was chosen since, by this time in the 10-week semester, students were familiar with their instructors' teaching behaviors and capable of rating instructors accurately. A

four-minute explanatory video was also posted as an announcement in the classes. Faculty were asked to play the video during week 8 seminars in these courses. Interested students clicked on an embedded link, taking them to a SurveyMonkey.com webpage. The SurveyMonkey landing page allowed students to read and agree to the Informed Consent.

Using this recruitment method, 32 unique faculty teaching 38 unique courses across seven academic departments and 213 unique students responded to the study (see Table 1). Mean student age was 36.1 years-old; most were women, 82.2%, with 17.8% men. These demographics are typical of online education (OE) and of PG's student body (Purdue Global Office of Reporting and Analysis, 2018; Seaman, Allen, & Seaman, 2018). Most respondents, 92.1%, were undergraduates, relatively evenly divided across the four years; see Table 1. Percentages of respondents by department also appear in Table 1.

Measures

Online Teaching Effectiveness Scale (OTES). The OTES contains closed-ended questions. The researchers developed the OTES through a thorough review of existing OTE research (Blackman et al., 2019) combined with their expertise as online instructors. Blackman et al. developed a precise definition of OTE. *Online teaching effectiveness* involves instructors facilitating student learning and construction of knowledge by:

1. *Presence*—strong cognitive, social, and teaching presence, promoting learning through social constructivism, effective communication, and quality instructional techniques.
2. *Engagement*—directly fostering engagement in the classroom, including timely and facilitative feedback and relationship building.
3. *Expertise*—demonstrating and applying content expertise and maintaining technical expertise.
4. *Facilitation*—regular, active, and thoughtful classroom interactions executing planned activities, managing communications, and supervising learning processes.

Using this definition, the researchers developed a set of pilot items to represent the OTE construct, with a minimum of five pilot items for each of the four dimensions. The pilot items can be found in Table 2.

Student Evaluation of Online Teaching Effectiveness (SEOTE). Bangert's (2006, 2008) 26-item SEOTE, a freely-available measure, strongly focuses on instructors' actual online teaching competencies and less on course characteristics beyond online instructors' control. Bangert developed his questions in relation to Chickering and Gamson's (1987) approach to OTE, which reflects constructivism. Thus, the theoretical foundation of the SEOTE accords with the primary theoretical approach to OTE in the literature. SEOTE items are scored using a six-point Likert scale (1 = strongly disagree; 6 = strongly agree). Bangert demonstrated SEOTE content validity through expert reviews by online instructors. Reliability of each factor was demonstrated via internal consistency estimates of $\geq$ .84. Although the SEOTE was published more than 10 years ago and does not reflect OTE research from recent years, the SEOTE remains among the best-validated extant OTE measures. As such, it served in the present study as a comparison measure to assist with OTES construct validation.

Outcome measures for construct validation. Two outcome measures were available: (a) participants' anticipated course grade in the course whose instructor they rated and (b) participants' actual grade in this course, issued by the instructor after the students' OTE rating.

Procedures

Participating instructors were notified by email exactly when to email the Research Announcement to all students in their classes, so that students received the Research Announcement during week 8 of the semester. Interested students clicked on the link in the Research Announcement. If comfortable with the Informed Consent, they indicated this electronically and completed the survey questions. The survey remained open for one week following the email send date.

To obtain test-retest data, the same students received a similar email one week later, during week 9 of the course. Students were asked to rate their instructors again via the same pool of OTE questions. Demographic and SEOTE questions were omitted, as a second set of responses to these items was not needed to establish test-retest reliability. Once the retest survey closed, outcome data were obtained from the university.

Statistical Analysis

Myers, Ahn, and Jin (2011) recommend a sample of $N \geq 200$ as an a priori target for sufficient statistical power to perform factor analysis. Thus, the 200 respondents in the present study sufficed for the analyses. A principal component analysis (PCA) reduced and refined the OTES to its principal components, which were based solely on extracted factors and not on a priori theory. Items comprising the principal components were then entered into a CFA with varimax rotation to verify factor structure, followed by Cronbach's alpha computations to measure internal consistency. The resulting OTES contained only items loading within confirmed factors.

The researchers reviewed the confirmed OTES factors for thematic unity and developed names to capture the items loading on each factor. The final retained OTES items were then compared within subjects to the same items for test/retest purposes, determining test/retest coefficients for each OTES factor and the total OTES score.

Initial construct validity for the OTES was established using a multitrait, multimethod approach to explore indicators of both convergent and discriminant validity (Campbell & Fiske, 1959). Convergent validity was explored by comparing the total OTES score to the "overall teaching effectiveness item" on the survey, as in Young's (2006) study. OTES total and factor scores were also correlated with SEOTE scores. Given that OTE is a measure of instructor performance, predictive validity was assessed by correlating mean OTES factor scores for all students taught by each instructor with mean course grades for that instructor. Higher instructor OTES scores were expected to be associated with higher grades, reflecting greater instructor teaching effectiveness. Discriminant validity was established by correlating OTES scores with student age and anticipated grade and contrasting via MANOVAs OTES scores by gender, year in school, and the department to which the course belonged. The OTES was not expected to be associated with these factors, as a robust measure of OTE should not be influenced by age, gender, student status, field of study, or anticipated course grade, but should produce similar results across these variables.

Results

Scale Construction

Data screening and sampling. The data were screened for univariate outliers. None was detected. Initial sample contained 269 respondents, but 56 surveys were incomplete and thus removed, leaving a final sample of 213.

A power analysis for Pearson correlations was conducted in G*Power to determine a sufficient sample size with an alpha of 0.05, a power of 0.80, and a large effect size, $\rho = .5$ (Faul, Erdfelder, Buchner, & Lang, 2013). Based on these assumptions, the desired sample size was 111 or greater. The current sample of 213 exceeded this minimum and permitted sufficient statistical power to detect moderate and large effect sizes.

Traditional recommendations (Catell, 1979) suggest a ratio of N: $p \approx 3$ to 6 to provide sufficient power for factor analysis. This requires a sample of 250. However, more recently, MacCallum, Widaman, Zhang, and Hong (1999) argued that this guideline is not mathematically sound. Osborne and Costello (2004) demonstrated that larger sample sizes open researchers to higher probabilities of error and component loadings that overlook latency. Ultimately, then, factor loading within a minimum number of iterations (< 100) demonstrated that the present sample provided sufficient statistical power for factor analysis (MacCallum et al., 1999). Furthermore, the Kaiser-Meyer-Olkin (KMO) sampling adequacy measure was .936. KMO values between 0.8 and 1.0 indicate adequate samples.

Testing assumptions and data pretreatment. PCA was chosen to avoid bias or attempts to fit data to proposed OTE factors. The goal, then, was to extract factors from data provided by student responses to items assessing their instructors' OTE (Tabachnick & Fidell, 2013).

Data pretreatment included performing Pearson correlations on all initial OTES items to avoid high collinearity that could produce false factors. The assumption was that if an item-item correlation was very high, the two items assessed the same part of the construct. Using an initial cut off of $r = .8$ with more than four other items, 10 items were removed, leaving 40 items for the PCA where the highest levels of collinearity were $r < .75$. Bartlett's test of sphericity was significant: χ^2 (780) = 10885.76, $p < .0001$.

Factor analysis and PCA. The remaining 40 items were entered into a PCA using SPSS version 25, after subjecting the data to a check for suitability for the analysis. The PCA revealed the presence of four components with eigenvalues exceeding 1, explaining 64.7%, 4.3%, 3.7% and 3.1% of the variance, respectively, with the four-component solution explaining 75.47% of the variance. An inspection of the scree plot revealed a clear break after the second component.

To aid interpretation of these four components, varimax rotation was performed. Within 20 iterations, the rotated solution revealed the presence of four clear factors showing several strong loadings and variables loading substantially on all four components. Items loaded on components as follows: Component 1, Presence, 21 items; Component 2, Expertise, seven items; Component 3, Facilitation, seven items; and Component 4, Engagement, five items. See Table 2 for the factor loadings of each item.

CFA. Prior to the CFA, the authors determined that, based on theory, a 40-item survey was too long. The goal of effective survey design is to measure constructs with short, concise, user-friendly questions that produce high response rates (Saleh & Bista, 2017). Scale purification was therefore conducted (Wieland, Durach, Kembro, & Treiblmaier, 2017). The authors decided to include all four factors with a minimum of two items per factor. This model included Presence with six items and the remaining three factors with two items apiece.

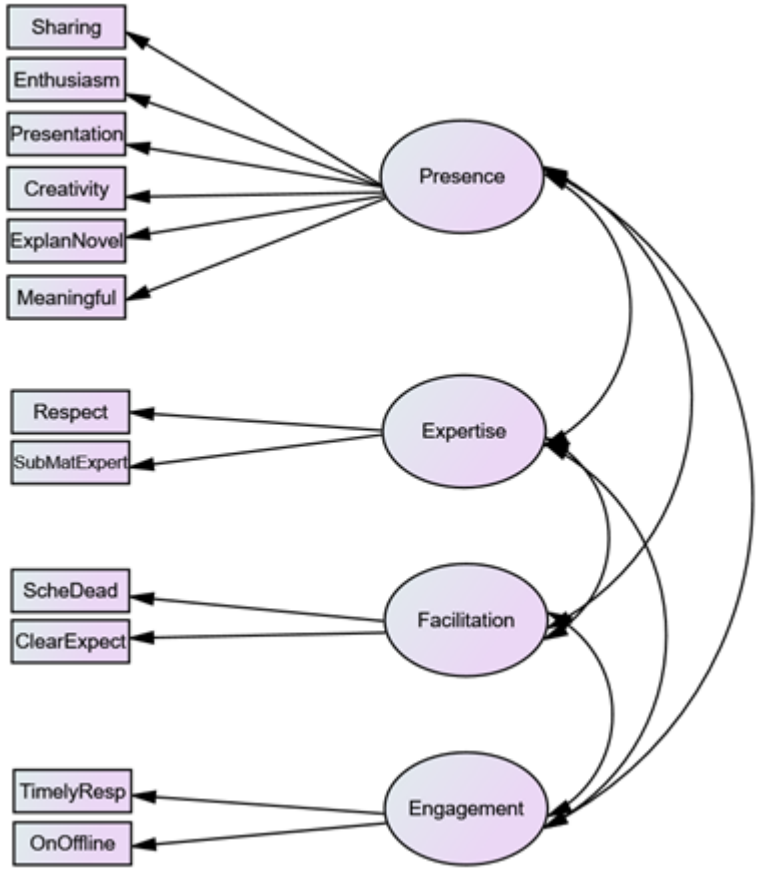

Figure 1. 12-item OTES

The CFA, using the estimation method of maximum likelihood over the variance-covariance matrix for the four-factor model, was conducted via the AMOS 25.0 statistical package (Arbuckle, 2014). To achieve model identification, regression coefficients of error terms over endogenous variables were fixed to 1. The CFA including goodness-of-fit modeling was performed to determine whether the PCA-derived four-factor model fit the actual data. Results indicated that the four-factor model fit the data well, χ^2 (48) = 255.41, p = .0001, RMSEA = 0.143, TLI = 0.857, CFI = .912. This model appears in Figure 1. Factor loadings for the final 12 items appear in Table 3.

Reliability

Cronbach's alphas for the four OTES factors and total scale were: Presence, .95; Expertise, .68; Facilitation, .81; Engagement, .82; Total, .95. Test/retest reliability coefficients for the four factors and total OTES scale ranged from r = .74 to .89; all were significant at p < .001, one-tailed. Coefficients were: 1. Presence, r = .85; Expertise, r = .74; Facilitation, r = .74; Engagement, r = .87; Total, r = .89.

Validity

Table 4 presents OTES inter-correlations and correlations with similar measures. The OTES total and all four factor scores correlated significantly, p < .001, with the overall teaching

effectiveness item, with coefficients ranging from r = .50 to .72. OTES total and factor score intercorrelations were all significant at $p < .001$, with coefficients between factors ranging from r = .49 to .71 and all factors having greater correlations with the total score than with other factors. OTES total and factor scores also correlated significantly, $p < .001$, with all four SEOTE scale scores, with coefficients ranging from r = .38 to .69. The lowest correlations with OTES scores occurred for the SEOTE Active Learning and Student Cooperation factors.

Correlations were run for the OTES total and factor scores with mean course grades. Course grade was significantly correlated with instructor Expertise measured by the OTES, r = .1, p = .05, one-tailed. Course grade was virtually uncorrelated with OTES Presence, Engagement, and Facilitation factors.

OTES total and factor scores were not significantly correlated with student age or anticipated grade, with correlations ranging from r = -.03 to .07 and none approaching significance. Three MANOVAs were computed to analyze OTES total and factor scores in relation to: (a) gender; (b) student status; and (c) department. For gender, the overall model was not significant: *Wilks' Lambda* = 0.967, $F(8,416) = 0.876$, $p = 0.54$. For student status, the overall model was not significant: *Wilks' Lambda* = 0.936, $F(24,709) = 0.565$, $p = 0.95$. For department, the overall model was not significant: *Wilks' Lambda* = 0.910, $F(28,730) = 0.689$, $p = 0.89$.

Discussion

Across the 30-year history of online education, the construct of OTE had not been well-defined or established through research. Rather, the face-to-face teaching paradigm was imported with little modification and applied to OTE, even though online teaching diverges from face-to-face teaching in substantive ways (Blackmann et al., 2019). In 2019, Blackmann et al. completed a comprehensive literature of OTE and developed a theoretical framework for conceptualizing the key dimensions of OTE. In the present research, we built upon Blackmann et al.'s theoretical framework and asked: 1. What factors empirically comprise the construct of OTE? 2. Does an empirically-derived OTE measure demonstrate adequate reliability and validity? The answer to both questions appears to be "yes."

For instance, the current results demonstrate that the theoretical framework for OTE emerging from Blackmann et al.'s recent literature review—a framework with the four OTE dimensions of Presence, Expertise, Facilitation, and Engagement—was supported by the quantitative analyses. When theoretical constructs align well with factors derived from purely quantitative analyses, as in the present research, the theoretical concepts can be taken to represent a strong framework for actual behavior in the specified domain. Thus, the match between theory and data in the present case supports the construct validity of the proposed OTE framework and the OTES. Construct validity of the OTES was also supported by supplemental analyses: total and factor scores were not significantly correlated with student age, gender, status, department, or anticipated grade—student-specific features that should not affect OTE, since OTE is a measure of instructor factors and not student factors. Furthermore, OTES reliability measures were relatively strong. Thus, the OTES appears to be a robust measure capable of assessing OTE among students from various backgrounds and education levels. This makes the OTES potentially applicable across diverse OE settings.

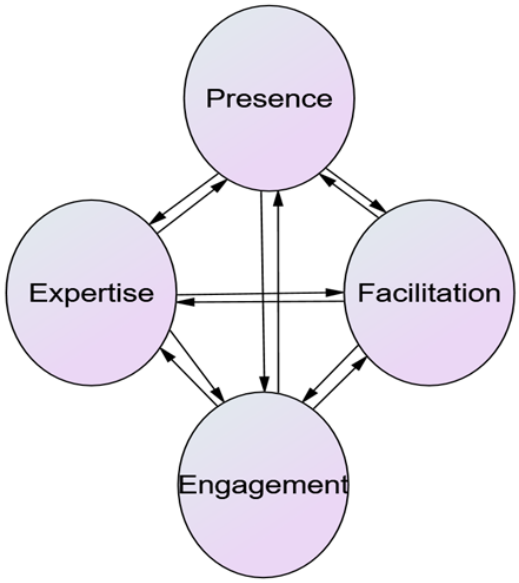

Figure 2. Synergistic Model of OTE

The OTES contains 12 items measuring four factors that replicate both theoretically and empirically: Presence, Expertise, Facilitation, and Engagement. Presence proved to be the most important OTE factor. Presence included items measuring instructors' ability to share relevant experiences and provide meaningful examples to illustrate course material, present information creatively, and communicate enthusiasm for course topics. Expertise included items measuring instructors' subject-matter knowledge and respect for students. Facilitation and Engagement factors were reminiscent of Chickering and Gamson's (1989) Seven Principles of Good Practice, since these factors included items on providing students with clear expectations, reminding them of deadlines, instructor availability, and instructor responsiveness. All four OTES factors were somewhat interrelated via CFA, as Figure 2 shows, and together provide a complete representation of the OTE construct. Thus, the total OTES score consists of the sum of the four factors and should succinctly and accurately indicate overall OTE.

Online Teaching Effectiveness Framework

The comparison of the OTES to the SEOTE (Bangert, 2006; 2008) revealed strong correlations between the two. Yet the OTES scores correlated more strongly with the OTES single-item overall teacher rating than with the SEOTE scales. This supports the validity of the OTES, in that the SEOTE measures teaching effectiveness for brick and mortar settings and includes questions beyond the scope of OTE, whereas the overall teacher rating is a single question about OTE itself.

In comparing the OTES and SEOTE, the lowest correlations occurred across all four OTES factors and the SEOTE Student Cooperation factor. This SEOTE factor is a direct reference to Chickering and Gamson's (1989) principle of Cooperation Among Students. It is logical that this factor was least correlated with the OTES because OE generally does not include a high degree of inter-student cooperation or many group projects. Due to the medium of instruction and that online students are often older working adults, the ability to cooperate with classmates is less important in OE than in traditional education settings serving younger students whose principal undertaking is often the pursuit of their university education (Lewis, 2016).

Limitations

Regardless of the mathematical assurance of a sufficient sample size in the present study, the development of the OTES would have benefited from a larger sample that included more graduate students. Further development of the measure could also benefit from student samples at other online schools. Although Purdue Global has a 20-year history of providing quality OE across a range of schools, colleges, and diverse academic and applied departments, it is only one institution and may not be representative of OE in general. The measure and OTE framework would also benefit from separate data collections to provide additional confirmation of the four OTE factors. The consensus is that when developing a measure, it is best to use different data sets or some form of cross-validation to perform a CFA (Cabrera-Nguyen, 2010). The current use of one data set for both PCA and CFA, although convenient, should be enhanced by further study to validate both the OTE framework and measure.

Conclusion

As a recent comprehensive literature review revealed (Blackmann et al., 2019), previous OTE measures, such as the SEOTE (Bangert 2006; 2008), were based on the Seven Principles of Good Practice, a theory created by college educators on ground campuses (Chickering & Gamson, 1989). This theory has been used for 30 years to measure teaching effectiveness at traditional universities. Applying this model to OE entails theoretical limitations, because OE differs substantially from ground education. For example, OE expands educational access to students who cannot reach ground campuses for various reasons, such as distance or disability. OE strongly supports students who have been poorly prepared for college. OE may in some cases provide a more rigorous education: for example, a student may receive a B- in a traditional course compared to a C in the same course offered online (Bettinger & Loeb, 2017). Attempting to mimic on-the-ground instruction does not account for the particular teaching skills needed to reach the diversity of students enrolled in online universities and negates the exclusive features and advantages of OE (Bettinger, Fox, Loeb, & Taylor, 2015). Thus, a new framework for OTE was needed to capture the unique dimensions of online teaching.

The OTES establishes this new framework for effective online teaching. The analyses herein support the theoretically-established understanding that online teaching indeed requires different instructor qualities than on-the-ground teaching. However, more research is required to refine and expand on the framework, in particular to validate the framework at additional online universities. Having a strong measure of OTE will assist online institutions in hiring and training instructors to provide more effective OE and engage in valuable quality improvement activities.

Author's Note

All authors originally affiliated with the Department of Graduate Psychology, Purdue University Global.

Corresponding author: Elizabeth Reyes-Fournier, elizabeth.reyesfournier@purdueglobal.edu.

The authors wish to express sincere appreciation to Dr. Sara Sander, Dean and Vice President of the College of Social and Behavioral Sciences at Purdue University Global, and to Dr. Julee Poole, the Chair of the Purdue University Global Graduate Psychology Program, for their continued support with this project.

References

Arbuckle, J. L. (2014). *Amos* (Version 23.0) [Computer software]. IBM SPSS. https://www.ibm.com/support/pages/what-correct-format-citing-amos

Bangert, A. W. (2006). The development of an instrument for assessing online teaching effectiveness. *Journal of Educational Computing Research, 35*(3), 227–244.

Bangert, A. W. (2008). The development and validation of the Student Evaluation of Online Teaching Effectiveness. *Computers in the Schools, 25*(1–2), 25–47. doi:10.1080/07380560802157717

Berk, R. A. (2013). Face-to-face versus online course evaluations: A "consumer's guide" to seven strategies. *Journal of Online Learning and Teaching, 9*(1), 140–148.

Bettinger, E., Fox, L., Loeb, S., & Taylor, E. (2015). *Changing distributions: How online college classes alter student and professor performance* (CEPA Working Paper No.15–10). Stanford Center for Policy Analysis. http://cepa.stanford.edu/wp15-10

Bettinger, E., & Loeb, S. (2017). Promises and pitfalls of online education. *Economic Studies at Brookings: Evidence Speaks Reports*, *2*(15).https://www.brookings.edu/wp-content/uploads/2017/06/ccf_20170609_loeb_evidence_speaks1.pdf

Blackman, G., Pedersen, J., March, M., Reyes-Fournier, E., & Cumella, E. J. (2019). A *comprehensive literature review of online teaching effectiveness: Reconstructing the conceptual framework* [Unpublished manuscript].

Cabrera-Nguyen, E. (2010). Author guidelines for reporting scale development and validation results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research, 1*, 99–103. doi:10.5243/jsswr.2010.8

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

Catell, R. B. (1979). *The scientific use of factor analysis*. Plenum.

Centra, J. A. (2005). *The development of the Student Instructional Report II*. Educational Testing Service. https://www.ets.org/Media/Products/283840.pdf

Chickering, A. W., & Gamson, Z. F. (1989). Seven principles for good practice in undergraduate education. *Biochemical Education***,** *17*(3), 140–141. doi:10.1016/0307-4412(89)90094-0

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2013). G*Power Version 3.1.7 [computer software]. Uiversität Kiel, Germany. Retrieved from http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register

Klieger, D., Centra, J., Young, J., Holtzman, S., & Kotloff, L. J. (2014). *Testing the invariance of interrater reliability between paper-based and online modalities of the SIR II™ Student Instructional Report*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/SIRII-Report-Klieger-Centra-2014.pdf

Lewis, M. (2016). Demographics of online students. In S. Danver (Ed.), *The SAGE encyclopedia of online education* (pp. 311–313). SAGE. doi: 10.4135/9781483318332.n103

Lokken, F. (2016). *ITC Annual National eLearning Report 2016 survey results*. https://associationdatabase.com/aws/ITCN/asset_manager/get_file/154447?ver=297

Liu, O. L. (2011). Student evaluation of instruction: In the new paradigm of distance education. *Research in Higher Education, 53*(4), 471–486. doi:10.1007/s11162-011-9236-1

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84–99.

Myers, N., Ahn, S., & Jin, Y. (2011). Sample size and power estimates for a confirmatory factor analytic model in exercise and sport: A Monte Carlo approach. *Research Quarterly for Exercise and Sport, 82*(3), 412–423. doi: 10.5641/027013611x13275191443621

National Center for Education Statistics. (2018). Table 311.15. Digest of Education Statistics (NCES 2017–094). https://nces.ed.gov/programs/digest/d16/tables/dt16_311.15.asp

Nilson, L. B., & Goodson, L. A. (2017). *Online teaching at its best: Merging instructional design with teaching and learning research* (1st ed.). Jossey-Bass.

Osborne, J. W., & Costello, A. B., (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research & Evaluation, 9*(11). http://PAREonline.net/getvn.asp?v=9&n=11

Pike, G. R. (2004). The Student Instructional Report for distance education: e-SIR II. *Assessment Update*, *16*(4), 11–12.

Purdue Global Office of Reporting and Analysis. (2018). *Purdue Global facts: World-class education online*. https://www.purdueglobal.edu/about/facts-processes/

Saleh, A., & Bista, K. (2017). Examining factors impacting online survey response rates in educational research: Perceptions of graduate students. *Journal of Multidisciplinary Evaluation*, *13*(29), 63–74.

Seaman, J. E., Allen, I. E., & Seaman, J. (2018). *Grade increase: Tracking distance education in the United States*. Babson Survey Research Group. http://www.onlinelearningsurvey.com/highered.html

Serdyukov, P. (2015). Does online education need a special pedagogy? *Journal of Computing & Information Technology, 23*(1), 61–74. doi: 10.2498/cit.1002511

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Pearson.

Thomas, J. E. & Graham, C. R. (2017). Common practices for evaluating post-secondary online instructors. *Online Journal of Distance Learning Administration, 20*(4). https://www.westga.edu/~distance/ojdla/winter204/thomas_graham204.html

Wieland, A., Durach, C. F., Kembro, J., & Treiblmaier, H. (2017). Statistical and judgmental criteria for scale purification. *Supply Chain Management: An International Journal, 22*(4), 321–328. doi: 10.1108/SCM-07-2016-0230

Young, S. (2006). Student views of effective online teaching in higher education. *American Journal of Distance Education, 20*(2), 65–77.

Appendix A: Tables

Table 1
Respondents' Demographic and Course Characteristics (N = 213)

Measure	All Subjects
Age	36.1 (9.7)
Gender	
Male	17.8%
Female	82.2%
School Status	
Freshman	24.4%
Sophomore	23.9%
Junior	21.6%
Senior	21.6%
Master's Program	7.0%
Graduate Certificate Program	0.9%
Department in Which Rated Course Was Taken	
Communication	25.0%
Criminal Justice	15.3%
Early Childhood Education	1.9%
Fire Science	6.3%
Human Services	8.2%
Psychology	41.8%
Social Science	0.5%
Number of Unique Faculty Rated by Respondents	32
Number of Unique Courses in Which Respondents Enrolled	38

Table 2
Principal Components Analysis Varimax Rotation with Kaiser Normalization

Items	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
COMPONENT: PRESENCE				
Sharing their relevant	.778	.299	.162	.314
Enthusiasm for teaching	.777	.173		.360
Good presentation skills	.765	.192	.396	.237
Creativity to increase	.760	.188	.361	.324
Explanations presentations	.745	.181	.495	.168
Meaningful examples	.735	.192	.320	.155
Explanation to improve	.691	.337	.249	.283
Facilitation of thoughtful	.659	.451	.358	.169
Effective communication	.658	.428	.159	.318
Getting to know students	.642	.244	.272	.393
Motivation to take responsibility	.633	.413	.407	.227
Effective facilitation	.626	.323	.415	.256
Explanations of complex	.608	.423	.438	.229
Giving students valuable	.600	.463	.322	.263
Personalized interactions	.588	.476		.516
Emphasis of important	.583	.408	.340	
Giving students clear	.556	.387	.414	.408
Encouragement to students	.547	.388	.409	.401
Action oriented feedback	.541	.321	.371	.389
Effort to create a comfortable	.541	.472	.211	.481
Efficiency	.513	.171	.478	.506
COMPONENT: EXPERTISE				
Respect for students	.127	.866	.272	.110
Subject matter knowledge	.428	.692	.104	.130
Tolerance	.356	.651		.487
Honesty and Integrity	.234	.641	.455	.304
Resourcefulness	.370	.600	.519	.235
Motivation for student	.433	.546	.271	.396
Concern about student	.435	.460	.392	.413

COMPONENT: FACILITATION				
Schedules and deadlines	.206	.270	.717	.260
Clear expectations	.341	.450	.675	.199
Encouragement for	.408	.236	.613	.377
Extra resources for learning	.574	.164	.589	.103
Professionalism	.229	.576	.583	.282
Organization	.487	.377	.575	.257
Helping students to	.539		.574	.374
COMPONENT: ENGAGEMENT				
Timely responses to	.308	.310	.202	.686
Online and offline availability	.309	.162	.359	.674
Timely grading of material			.511	.601
Understanding and as	.300	.347	.306	.572
Warmth and friendliness	.519	.316	.130	.540

Table 3
Factor Loadings of Final 12-Item OTES

	Factors			
Presence	1	2	3	4
Sharing their relevant professional experiences	.778	.299	.162	.314
Enthusiasm for teaching	.777	.173		.360
Good presentation skills	.765	.192	.396	.237
Creativity to increase student interest	.760	.188	.361	.324
Explanations/presentations of material in novel ways	.745	.181	.495	.168
Meaningful examples	.735	.192	.320	.155
Expertise				
Respect for students	.127	.866	.272	.110
Subject matter knowledge	.428	.692	.104	.130
Facilitation				
Schedules and deadlines	.206	.270	.717	.260
Clear expectations	.341	.450	.675	.199
Engagement				
Timely responses to questions	.308	.310	.202	.686
Online and offline availability	.309	.162	.359	.674

Table 4
OTES Inter-Correlations and Correlations with Similar Measures

	OTES				
Measure	Presence	Expertise	Facilitation	Engagement	Total
Overall Rating	.71	.50	.60	.53	.72
OTES Presence	.60				
OTES Expertise	.71				
OTES Facilitation	.67	.62			
OTES Engagement	.96	.49	.59		
		.72	.83	.79	
Overall Rating	.82	.60	.70	.62	.84
SEOTE Student/Faculty Inter.	.62	.49	.59	.64	.69
SEOTE Active Learning	.58	.38	.57	.50	.61
SEOTE Time on Task	.67	.46	.59	.53	.69
SEOTE Student Cooperation	.51	.42	.46	.40	.53

Note. All correlation coefficients were significant at $p < .001$, one-tailed.