An integrated test of multidimensionality, convergent, discriminant, and criterion validity of the Course Experience Questionnaire: an Exploratory Structural Equation Modeling

Short title: *The Course Experience Questionnaire*

Francisco Cano*
Faculty of Psychology.
University of Granada, Spain

M.C. Pichardo; A.B.G. Berbén; M. Fernández-Cabezas
Faculty of Educational Sciences
University of Granada, Spain

*Corresponding author information:
Francisco Cano. *Department of Educational Psychology. Faculty of Psychology. University of Granada -Spain-. Campus de Cartuja s/n, 18071 Granada (Spain).* Phone : +34 958 24 06 74, Fax : +34 958 24 90 17, E-mail: fcano@ugr.es.

Acknowledgements
The authors wish to thank Professor George Franke for his advice and support in applying the Mplus codes for inferential tests of the HTMT ratios.

An integrated test of multidimensionality, convergent, discriminant, and criterion validity of the Course Experience Questionnaire: An Exploratory Structural Equation Modeling

Abstract

Most research on the course experience questionnaire (CEQ) has been conducted through conventional exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) within the independent cluster model framework (ICM-CFA). However, very few studies have focused on examining its multidimensionality using more flexible psychometric frameworks such as exploratory structural equation modelling (ESEM).

This study aims to conduct an integrated test of multidimensionality on the short, 23-item version of the CEQ (CEQ23) by using ESEM, to test its construct and criterion-related validity and contribute to the current debate on its validity. The participants comprised 620 undergraduate psychology students. CEQ23 scores were examined through ESEM to identify two sources of construct-multidimensionality. This entailed contrasting ICM-CFA and ESEM solutions and comparing three alternative models. Construct and criterion-related validity were then analysed using common and emerging techniques.

The results (a) confirmed the presence of a superior construct: students' perceptions of teaching quality, which is multifaceted and hierarchically structured; (b) supported a generally acceptable construct and criterion-related validity; and (c) highlighted some methodological weaknesses of conventional statistical techniques, which may underly the debate on the validity of the CEQ23.

Keywords:

exploratory structural equation modeling (ESEM); perceptions of teaching quality; Course Experience Questionnaire (CEQ); psychometric multidimensionality; construct validity

Introduction

For decades, students' evaluation of teaching (SET) has been the prevailing way to assess teaching quality at universities worldwide, and more recently, has been used as a tool for quality assurance and accountability in higher education institutions (Spooren, Vandermoere, Vanderstraeten, and Pepermans 2017). Research in SET has demonstrated a trend of continuous growth, partly due to the complexity of teaching, a multidimensional construct involving many interrelated dimensions (e.g. clarity, organisation) with some conceptual overlapping (Marsh et al. 2009). This has given rise to the development of many measures or performance indicators (Ramsden, 1991; Slade et al. 2014) of which the Course Experience Questionnaire (CEQ) and, more specifically, its short 23-item version (CEQ23) is the most widely used (Richardson, 2009; Wilson, Lizzio, and Ramsden 1997). Nevertheless, methodological concerns have been raised about both the nature and dimensions of SET-measures and the results of validation procedures, which 'calls into question the structural validity of these instruments' (Spooren et al. 2017, 238).

Although much research has examined the construct validity of the CEQ (see Richardson's 2009 review), relatively little has focused on its construct-relevant multidimensionality. Both conventional Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) have been widely used within the independent cluster model framework (ICM-CFA) (Richardson, 1994; Wilson et al. 1997). However, these statistical techniques are not without their weaknesses (e.g. over-estimation of inter-factor correlations) (Asparouhov and Muthén 2009), which negatively affect the results obtained. This has motivated the design of more flexible approaches, such as exploratory structural equation modeling (ESEM; Asparouhov and Muthén 2009; Marsh et al. 2009), which integrates EFA and ICM-CFA, allows the testing of a priori hypotheses about structural validity, and is a crucial element of Morin, Arens, and Marsh's (2016) guidelines for conducting a comprehensive test of construct-multidimensionality.

Based on the above, the aims of this study are to conduct an integrated test of multidimensionality of the CEQ23 using ESEM, to test its construct and criterion-related validity, and thus to contribute to the debate about its validity for assessing teaching quality in higher education.

Below, we offer a short review of i) the development and psychometric properties of the CEQ; ii) studies that have examined its multidimensional structure; and iii) the advantages of ESEM over previous statistical techniques and Morin et al.'s (2016) integrated test of multidimensionality.

Development of the CEQ

The theoretical and empirical framework that underlies the CEQ can be traced back to research generated at Lancaster University in the 1980s, which focused on students' perceptions of the teaching-learning environment as determinant factors of their learning approaches and subsequent learning outcomes (e.g. Ramsden and Entwistle 1981). Two important milestones in devising the CEQ were Elphinstone's (1989) Master's thesis, from which a first version of the CEQ (comprising 46 items in 8 factors) was derived, and Ramsden's (1991) national trial in Australia of a revised version comprising 30 items in five scales (e.g. Good teaching, Emphasis on independence).

Although the CEQ was initially designed to assess graduate students' perceptions of teaching quality across their degree programmes, it has also been used with enrolled undergraduate students (e.g. Ginns, Prosser, and Barrie 2007) and even to rate particular course units (e.g. Kreber, 2003). These variations may be associated to some extent with the increasing demands experienced since the end of the 1990s (e.g. for quality assurance, employability of graduates, need for shorter and revised SET-instruments for quick and easy scanning), which gave rise to the CEQ23 (Byrne and Flood 2003; Wilson et al. 1997). This instrument i) omitted the Emphasis on independence scale, because its items cross-loaded on several factors; ii) included a new scale of Generic Skills; and iii) consisted of 23 items (24 in some studies, e.g. Curtis 2005) in five scales (Generic skills; Appropriate assessment; Appropriate workload; Good teaching; and Clear goals and standards) and one additional item for assessing students' overall satisfaction (see Richardson 2009; Slade et al. 2014; Wilson et al. 1997, for detailed reviews).

Psychometric testing of the CEQ

The psychometric properties of the CEQ have been extensively assessed through different statistical techniques. Rasch analyses (e.g. McInnis, Griffin, James, and Coates 2001; Waugh 1988) and ESEM (Marsh, Ginns, Morin, and Nagengast 2011) have been used by a handful of researchers, while EFA and ICM-CFA have been used by most.

Analyses of the long versions of the CEQ have generally yielded satisfactory results (e.g. internal consistency, validity) (e.g. McInnis et al. 2001; Richardson 1994, 2009; Ramsden 1991; Wilson et al. 1997). Analyses of the short versions of the CEQ through EFA have in several cases (e.g. Broomfield and Bligh 1998; Kreber 2003) detected six factors, a consequence of splitting the Good teaching dimension into two factors. In most cases, however, i) the multidimensional five-factor structure emerged through EFA (Byrne and Flood 2003; Espeland and Indrehus 2003), ICM-CFA (Jansen, van der Meer, and Fokkens-Bruinsma 2013; Ginns et al. 2007), both EFA and ICM-CFA (Curtis 2005; Wilson et al. 1997), and ESEM (Marsh et al. 2011); ii) the levels of internal consistency were acceptable (e.g. Jansen et al. 2013); iii) inter-scale correlation was moderate and positive (e.g. Byrne and Flood 2003); and iv) criterion validity was established, with scores on the CEQ showing positive links with outcomes such as Satisfaction, Generic skills, and Achievement (e.g. Wilson et al. 1997).

There is an ongoing discussion about which might be the best model, e.g. first-order (correlated) factors, or higher-order, to represent the underlying multidimensional structure of the CEQ. Although some studies suggest that the modal solution could be a single higher-order factor which would account for the variance in the first-order factors, named ‘monarchical hierarchy’ by Richardson (1994), a focus on the short versions of the CEQ reveals some controversial results. Thus, Wilson et al. (1997), who used a higher-order path analysis, suggested a two higher-order factor structure, whereas Byrne and Flood (2003), who utilised an EFA, detected a single higher-order factor. By contrast, after conducting an ICM-CFA, Curtis (2005) suggested a nested (i.e. bifactor) structure. However, it seemed that neither the shortcomings of bifactor models (e.g. a tendency to ‘overfitting’, which is in their favour) (Watts, Poorer, and Waldman 2019) nor some alternative bifactor model-derived indices recently proposed by Bonifay, Lane, and Reise (2016) were considered.

Most research in SET-instruments has been conducted through conventional EFA and ICM-CFA (e.g. Jansen et al. 2013; Richardson 1994; Wilson et al. 1997), which may have yielded biased results. As noted by Marsh et al. (2009) in the first substantive ESEM study on student evaluations of teaching and commented on by Morin et al. (2016), the reasons are i) that these statistical techniques have serious weaknesses (e.g. ICM-CFA fails to include non-zero cross-loadings and inflates inter-factor correlations), which usually leads to distortions in factors, in their correlations, and in structural relations with other constructs; ii) they may not be a suitable choice for assessing multidimensional constructs; and iii) they have been superseded by more flexible psychometric frameworks such as ESEM (Morin et al. 2016).

ESEM

ESEM overcomes the aforementioned weaknesses by integrating EFA and CFA and testing a priori hypotheses about construct-relevant multidimensionality, while allowing the presence of cross-loadings (Asparouhov and Muthén 2009; Marsh, Morin, Parker and Kaur 2014). A recent method called the ESEM-within-CFA (EwC) allows researchers to 'start with an ESEM model, and re-express it in the CFA framework using the start values generated from the initial ESEM model' (Morin and Asparouhov 2018, 1).

ESEM is most suitable for analysing SET-instruments because, as with teaching, they must reflect two sources of construct-relevant psychometric multidimensionality (Marsh et al. 2009; Morin et al. 2016): i) the fallible nature of indicators, as items may be related to one or more of the dimensions of teaching; and ii) the hierarchical nature of SET-instruments where items may show loadings on their own dimensions as well as (directly or indirectly) on a hierarchical construct. The first source of multidimensionality is determined by comparing CFA and ESEM solutions, whereas the second source is detected by comparing three models: first-order, hierarchical, and bifactor (Arens and Morin 2017; Morin et al. 2016). These have also been labelled, for the sake of clarity, as first-order or correlated factors, higher-order (indirect hierarchical), and bifactor (direct hierarchical/nested factors) (e.g. Canivez 2016).

However, there is a scarcity of research on the multidimensionality of the CEQ23 through ESEM. To the best of our knowledge, Marsh et al. (2011) is the only research team to have considered this possibility, but they did not fully assess the multidimensionality due to hierarchically-structured constructs and their results were published as supplemental analyses only. They used ESEM to test the a priori hypothesis of a five first-order (correlated) factor structure and detected the fallible nature of its indicators, finding that: i) both ESEM and ICM-CFA identified the five posited factors; ii) inter-factor correlations were larger for the ICM-CFA than for the ESEM solution; iii) the goodness-of-fit for the ESEM solution was better than that of ICM-CFA; and iv) a single higher-order factor emerged, which may be used as a measure of the perception of course quality.

Aims of the Present Study

This study aims i) to conduct an integrated test of multidimensionality of the answers to the CEQ23 at item level by using ESEM in line with Morin et al.'s guidelines (2016), which involves contrasting ICM-CFA and ESEM solutions and comparing three alternative models (first-order, single higher-order, and bifactor); ii) to test convergent, discriminant, and criterion-related validity; and thus iii) to contribute to the current debate about the validity of CEQ23 scores for assessing teaching quality in higher education.

The following hypotheses are considered:

a) ESEM models will fit the data substantially better than their corresponding ICM-CFA models.

b) Inter-factor correlations will be substantially lower for ESEM than for ICM-CFA models.

c) Of the three models examined, the ESEM single higher-order model will be the best representation of the multidimensional structure of the CEQ responses.

d) Convergent, discriminant, and criterion validity will be generally acceptable.

Method

Participants and Procedure

A total of 620 undergraduate psychology students in the last two years of their course participated voluntarily in the study during the final teaching week of the spring semester. The majority were female (79.2%), enrolled in their penultimate year (54.2%), and aged between 19 and 25 (95.9%). Although participants were asked to provide their age, sex, and academic year, they all completed the questionnaire anonymously during regular class time.

Measures

Course experience

Course experience was assessed using the 23-item version of the CEQ (Wilson et al. 1997), grouped into five scales measuring *Appropriate assessment* (AAS) (3 items) (e.g. *To do well on this course all you really need is a good memory,* reversed item); *Appropriate workload* (AWS) (4 items) (e.g. *We are generally given enough time to understand the things we have to learn*); *Good teaching* (GTS) (6 items) (e.g. *This course really tries to get the best out of all its students*); *Generic skills* (GSS) (6 items) (e.g. *This course has helped me develop the ability to plan my own work*), and *Clear goals and standards* (CGS) (4 items) (e.g. *It's often hard to discover what's expected of you on this course,* reversed item). For validation purposes, the CEQ includes a single-item assessing students' overall satisfaction with their course quality (OSI). Items on these measures were rated on a 5-point Likert-type scale ranging from 1 ('*definitely disagree*') to 5 ('*definitely agree*').

Results

Initial analyses

Outliers were screened using z-scores on the 23 variables and then recoded to preserve as much data as possible. Univariate outliers (i.e. z scores ± 3.29) were recoded to the next highest or lowest value within the normal distribution (Tabachnick and Fidell 2001). One multivariate outlier was identified by means of the function outproad in R (Wilcox, 2005) at the rate of .05, declared missing, and imputed with a weighted-median through the k-nearest neighbours method.

The assumption of univariate (and multivariate) normality was not met because: i) univariate analyses showed that median values for kurtosis and skewness for the 23 variables were -.91 (range -.30 to -.93) and -.09 (range .62 to -.62), respectively; ii) the Shapiro-Wilks's W statistic was significant (p <.001) for each variable, indicating high kurtosis; and iii) the results of Mardia's tests of multivariate skewness and kurtosis were 11.57 (p <.001) and 4012.98 (p = 1.0), respectively.

Because of these conditions (non-normally distributed variables, with few response options and ordered-categorical, i.e. ordinal), polychoric correlations and a robust weighted least squares (WLSMV) estimator for estimation of factor structure in Mplus (Muthén and Kaplan 1992) were used. The Kaiser–

Meyer–Olkin measure of sampling adequacy exceeded 0.80 (MSA = 0.87), justifying the use of factor analysis.

Analyses of the sources of construct-relevant multidimensionality

These analyses were conducted to detect the existence of multidimensionality referring to: a) the fallible nature of indicators and the presence of conceptually related constructs, and b) the presence of hierarchical superior constructs.

Three representations of the underlying structure of the CEQ23 were estimated using conventional ICM-CFA as well as ESEM and then compared. The three conventional ICM-CFAs were: i) first-order (ICM-CFA), which included five (correlated) factors corresponding to the five CEQ23 subscales, each item loaded on its specific subscale with no cross-loadings between items and non-target factors allowed; ii) single higher-order (H-CFA), which was similar to the former, but adding for each first-order factor a loading on a higher-order factor; and iii) a bifactor CFA (B-CFA) in which all items simultaneously loaded on their respective specific first-order factors (set to be orthogonal) as well as on a general factor. The same series of representations was also estimated using ESEM in a confirmatory manner. These were: i) first-order ESEM (using target oblique rotation, main loadings, and cross-loadings ‘targeted’ to be near zero); ii) single higher-order hierarchical EwC (H-EwC) derived from the previous model; and iii) bifactor EwC (B-EwC) (with orthogonal bifactor target rotation) (Morin and Asparouhov 2018; Morin et al. 2016).

Contrasting conventional ICM-CFA and ESEM: the assessment of conceptually linked constructs. An examination of the goodness-of-fit statistics (see Table 1) revealed that regarding the conventional models, the ICM-CFA obtained lower goodness-of-fit indices (CFI = .884, TLI = .867, RMSEA = .070) than the corresponding H-CFA (CFI = .885, TLI = .871, RMSEA = .069), which, in turn, presented lower goodness-of-fit indices than the B-CFA(CFI = .927, TLI = .911, RMSEA = .057).

Table 1 about here

A similar pattern was found for the ESEM models. The fit of the H-EwC (CFI = .961, TLI = .936, RMSEA = .049) was higher than the ESEM, but lower than the B-EwC (CFI = .979, TLI = .959, RMSEA = .039). These results indicated that conventional ICM-CFA and H-CFA models failed to provide satisfactory goodness-of-fit indices, while overall these were considerably higher for all the ESEM models.

As expected, the comparison of the inter-factor correlations (see Table 2) showed that these were reduced substantially more for ESEM (|r| = .059 to r = .541, M = .292) than for ICM-CFA models (|r| = .165 to r = .737, M = .468).

Table 2 about here

These inflated inter-factor correlations of the ICM-CFA models somehow undermine i) the differentiation between the dimensions defining the factor structure of the CEQ, ii) its discriminant validity, and iii) its usefulness as a source of feedback to assess and improve those dimensions or components that the CEQ is designed to measure (Marsh et al. 2009).

By contrast, the inter-factor correlations from the best fitting models, i.e. ESEM, ranged from .059 (no correlation) to .541 (moderate correlation). This decrease was particularly marked in the case of F1-F3 factors, whose correlation decreased from r = .681 in ICM-CFA to r =.336 in ESEM; similar values were shown when comparing H-CFA and H-EwC models. These results suggest that models enabling cross-loadings (i.e. ESEM) i) tend to generate inter-factor correlations that are more representative of the actual value (Asparouhov and Muthén 2009), ii) support the conceptually different nature of the factors and consequently the construct-relevant multidimensionality of the CEQ23; and iii) should therefore be retained.

An examination of the parameter estimates of ESEM and H-EwC models revealed very similar factor loadings and cross-loadings inasmuch as the first-order structure of the latter corresponded to the measurement model of the former. Therefore, for the sake of simplicity and clarity, the estimates of ESEM were not reported. Parameter estimates corresponding to the H-EwC model (see the left-hand side of Table 3) suggested well-defined factors resulting from significant and substantial target factor loadings, which ranged from |λ| = .347 to .813, M = .585, and the majority (95.65%) were >.40.

Table 3 about here

When non-target cross-loadings were considered, 12 of the 92 possible ones were significant but low level (>.10 and <.20), while 8 were significant and substantial (>.20). The latter included, for example, items GTS16 (.260), GTS15 (.237), CGS23 (.224), and AWS14, which showed the highest cross-loading (.41), with similar factor loadings on the AWS and GTS factors. Most of the non-target cross-loadings involved constructs sharing some level of conceptual overlap, particularly between Good teaching and Appropriate assessment, and Appropriate workload and Clear goals, respectively. These results suggest: i) that the ICM-CFA assumption of including main loadings with no cross-loadings is highly restrictive and unreasonable for a multifactor instrument such as CEQ23; and ii) that evidence of construct-multidimensionality emerged in answers with this instrument because indicators are fallible in nature and assess conceptually related constructs. This possible common source of variation supports the appropriateness of relying on ESEM and of examining the hierarchical nature of the CEQ23.

Contrasting conventional ESEM AND B-ESEM: the presence of hierarchical superior constructs. The H-EwC solution, which, as previously noted, revealed well-defined factors, also provided evidence that: i) all but one of its first-order factors (Appropriate workload) loaded higher than .40 on the higher-order factor; and ii) did not lead to a significant decrease in model fit in comparison with the ESEM solution. Hence, it can be concluded that the H-EwC model provided a good representation of the construct assessed by the CEQ23.

The B-EwC provides an alternative representation of the CEQ23, in which multidimensionality is simultaneously viewed as a general factor (G-factor) and specific factors (S-factors) are assumed to be orthogonal. This model obtained the best goodness-of-fit indices and appeared to show acceptable parameter estimates (see the right-hand side of Table 3). A careful scan revealed, however, that neither was the G-factor sufficiently well-defined, with loadings ranging from $|\lambda|$ = .090 to .644 (M = .432), nor were the S-factors, with 12 S-factor target loadings higher and 11 lower than their corresponding G-factor loadings and the Appropriate workload S-factor being represented by just one loading >.30 (item AWS14). Moreover, the weaknesses of bifactor models (e.g. vulnerability to over-fitting data; tendency to yield false-positive results) (Watts et al. 2019) made it suitable for the calculation of some bifactor model-derived indices recommended by Bonifay et al. (2016). Thus, a) the Individual Explained Common Variance (IECV) yielded coefficients above .50 for only 39.13% of the 23 items; b) the ωH statistic was .726, lower than the 75% threshold; and c) the proportion of explained common variance (ECV) was only .435. These results indicate, respectively, that i) the bulk of the items were better measures of their S-factors than of the G-factor; ii) negligible unique variance was attributable to the G-factor; and iii) the strength of the G-factor was low compared to the S-factors. This suggests that the B-EwC model should be dismissed.

Taken together, the results indicate, in line with our hypotheses, the superiority of the H-EwC model over the other models, and confirm the presence of a superior construct, students' perceptions of teaching quality, which is multifaceted and hierarchically (indirectly) structured.

Analyses of convergent, discriminant, and criterion validity

Convergent validity was evaluated according to the usual three criteria (Bagozzi and Heatherton 1994): a) the magnitude of most factor loadings (threshold .50) was fair, as mentioned above; b) Average variance extracted (AVE) (threshold .50) was below the minimum acceptable except for Clear goals (see right-hand part of Table 4); and c) the composite reliability (ρ_C) (threshold .60 - .70) was acceptable for both specific factors (e.g. ρ_C for Good teaching = .766) (see left-hand part of Table 4) and the high-order factor (ρ_C = .934).

Table 4 about here

Discriminant validity was evaluated according to Fornell and Larcker's (1981) criteria (see left-hand part of Table 4). These criteria were met because: a) the square roots of AVE for any given pair of constructs were greater than the inter-construct correlations; b) the maximum shared variance (MSV, i.e. the square of the highest correlation coefficient, .418^2 = .175) was lower than AVE (see right-hand part of

Table 4); and c) the average shared variance (ASV, i.e. the mean of the squared correlation coefficients between latent constructs, .0096) was lower than AVE.

In addition, a recent estimate, more accurate than these criteria, of the true correlation between the two constructs was also applied, the heterotrait–monotrait ratio (HTMT, Henseler, Ringle, and Sarstedt 2015). This is defined as the mean value of indicator correlations across constructs (heterotrait) relative to the geometric mean of average correlations among indicators within the same construct (monotrait). We used constrained parameters in Mplus (see Franke and Sarstedt 2019) to calculate the HTMT ratios (and standard errors) and test them against a strict threshold value of .85. These ratios for each pair of constructs ranged from .153 to .709 (see Table 5) and were lower than .85, suggesting that discriminant validity was established (Henseler et al. 2015).

Table 5 about here

Criterion validity was proved with students' factor scores on the CEQ23, showing positive correlations with two learning outcomes measures (see Table 6).

Table 6 about here

The highest correlation was between course satisfaction and the higher-order factor (r = .530, p < .001), whereas the weakest was between Generic skills and Appropriate workload.

Discussion

The study aimed to improve our understanding of the measurement structure of the CEQ23 responses by providing an integrated test of its multidimensionality through the application of ESEM (Marsh et al. 2014; Morin et al. 2016) and to examine its construct validity and criterion-related validity.

First, generally consistent with our first and second hypotheses, the ESEM models yielded better goodness-of-fit indices and lower inter-factor correlations than their ICM-CFA counterparts, the latter indicating greater factor distinctiveness. Moreover, parameter estimates suggested well-defined factors (e.g.

those of the H-EwC model) according to the expected pattern of target factor loadings and non-target cross-loadings.

The presence of multiple, adequate low-level cross-loadings confirms previous research findings (e.g. Kreber 2003; Wilson et al. 1997) and is consistent with the assumption that the CEQ23 consists of separate dimensions sharing an acceptable conceptual overlap (Marsh et al. 2011). The presence of some substantial cross-loadings supports previous findings obtained i) through EFA, which indicated cross-loadings of item AWS14 in the GTS (Wilson et al. 1997), of item GTS16 on the GSS (Kreber 2003), and of item CGS23 on the GTS (Wilson et al. 1997), and ii) through ESEM, which indicated cross-loadings of items GTS15 and CGS23 on CGS (Marsh et al. 2011). These cross-loadings, particularly that of the strongest item AWS14, might suggest that the CEQ23 needs to be reviewed and updated, as proposed by Slade et al. (2014).

All these results are in line with i) Marsh et al.'s (2011) findings on comparing CFA and ESEM models of the CEQ23; ii) the general findings of the literature (e.g. Arens and Morin 2017; Asparouhov and Muthén 2009; Marsh et al. 2009; Morin et al. 2016), which have supported the superiority of ESEM solutions; iii) the detection of the source of multidimensionality concerning the fallible nature of items as indicators of a single construct and the finding of conceptually related constructs; and iv) Slade et al.'s (2014) proposal of continued revision and robust scale development of this SET-instrument.

Second**,** in agreement with our third hypothesis, the EwC provided the best representation of the CEQ23. This finding i) confirms the identification of the source of multidimensionality referring to the presence of hierarchical superior constructs, and ii) supports the idea that the five first-order latent factors of the CEQ23 are related to a multifaceted and hierarchically (indirect) construct, which accounts for their inter-correlations and can be interpreted as an index or global measure of perceptions of teaching quality. This conclusion on model selection is based not only on goodness-of-fit indices but also on an examination of parameter estimates and theoretical interpretability (Morin et al. 2016). Thus, the H-EwC solution yielded better goodness-of-fit statistics than the ESEM and had more well-defined and interpretable factors than the

B-EwC, which although seeming to obtain the best goodness-of-fit indices, failed to reach the significance threshold for model-derived indices recommended by Bonifay et al. (2016).

Our findings on the multifaceted and hierarchical structure of the CEQ23 are congruent with the results of the very few studies on this subject conducted through EFA (e.g. Byrne and Flood 2003; Richardson 1994) and ESEM (e.g. Marsh et al. 2011) and also with previous substantive theory (Richardson 2009). However, our findings are unique in the following ways: i) over and above the pure EFA, we applied an ESEM approach, which integrates many of the advantages of EFA and CFA (e.g. less biased parameter estimates, greater factor distinctiveness) (e.g. Asparouhov and Muthén 2009); ii) unlike Marsh et al. (2011), we contrasted first-order with hierarchical and bifactor models; and iii) to the best of our knowledge, this is the first time an integrated test of multidimensionality following Morin et al.'s guidelines (2016) has been conducted on the CEQ23. Interestingly, our results differ from those reported by Curtis (2005), who used CFA and suggested a nested (i.e. bifactor) structure for the CEQ23. There are three potential factors that might partly account for this difference: i) the possible weaknesses (e.g. substantially higher inter-factor correlations) induced by the CFA; ii) the exclusion of several bifactor model-derived indices (e.g. IECV, ECV) recently proposed by Bonifay et al. (2016); and iii) the possible lack of consideration of the statistical bias favouring the fit of bifactor models, which led Watts et al. (2019, 1287) to affirm that 'model fit statistics are unreliable indicators of the validity of bifactor models'.

Taken together, these findings i) demonstrate the presence of two sources of construct-relevant psychometrics in answers to the CEQ23; ii) corroborate that the higher-order (indirect hierarchical) model is the best way to represent a comprehensive measure of this SET-instrument, at least in the data considered here; and iii) indicate that some of Spooren et al.'s (2017) concerns about the structural validity of SET-forms may be due to the methodological weaknesses inherent in the conventional statistical techniques used and, in some cases, the almost exclusive use of goodness-of-fit indices when choosing between alternative models.

Third, in accordance with Hypothesis 4, our results substantiated previous findings in the literature (e.g. Byrne and Flood 2003; Marsh et al. 2011; Wilson et al. 1997), corroborating the generally acceptable convergent, discriminant, and criterion validity of CEQ23. In contrast with previous research, a new statistical tool, ESEM, and a rigorous assessment of discriminant validity (AVE and HTMT) were used in this paper.

This study has a number of limitations. First, although the sample size was sufficient to apply the ESEM approach, it only included participants from the psychology department within a particular university. In future studies, the sample should be generalised to include a larger number of participants and disciplines. Second, some substantial cross-loadings and relatively low convergent validity values were observed. Therefore, future research could revise the CEQ23 and (considering that it has not been revised since McInnis's (2001) work), update it to account for changing educational environments (e.g. the breakthrough of digital technologies in higher education) (Slade et al. 2014), without neglecting the importance of theoretical grounding and an appropriate research focus on student experience.

Notwithstanding these limitations, some methodological, theoretical, and practical implications can be derived from this investigation. Methodologically, this study illustrates in a simplified, step-by-step way the use of the relatively new ESEM framework for conducting an integrated test of the psychometric multidimensionality present in some SET-instruments. Most importantly, our findings i) challenge conventional EFA and ICM-CFA approaches to testing the nature of students' perceptions of teaching quality, as reflected in their answers to the CEQ23, by demonstrating that ESEM is a more appropriate and flexible alternative psychometric framework; and ii) warn of the risks of choosing among alternative models solely on goodness-of-fit indices. Theoretically, the modelling of perceptions of teaching quality as an integrated higher-order factor involves i) representing them by their shared variance across the first-order factors (i.e. dimensions); and ii) conceptualising them at a global level rather than at the level of those specific dimensions, which may promote a broader understanding and more holistic knowledge of this multifaceted construct, and could enhance parsimony and utility for policy and practice. Practically, researchers in SET-instruments in general and the CEQ in particular are encouraged i) to apply the ESEM integrative psychometric framework more extensively and systematically to address substantive critical issues about the underlying factor structure of many multidimensional instruments, ii) to improve their structural validity and thereby, iii) to facilitate the obtaining of valuable information on how teaching quality is perceived and the monitoring and improvement of teaching quality in higher education.

Conclusion

In summary, this study contributes to existing knowledge by a) demonstrating, through an ESEM integrative psychometric framework, the presence of two sources of construct-relevant multidimensionality in the CEQ23 and the detection of an ESEM single higher-order (indirect hierarchical) model as the best representation of the superior, multifaceted, and hierarchically structured construct of students' perceptions of teaching quality; b) highlighting several methodological weaknesses in conventional EFA and ICM-CFA, which may partly underlie the debate on the validity of SET-instruments in general and the CEQ more specifically, outlined by Spooren et al. (2017); and c) providing support for a generally acceptable construct validity (convergent and discriminant) and criterion-related validity.

References

Arens, A. K., and A. J. Morin. 2017. "Improved representation of the self-perception profile for children through bifactor exploratory structural equation modeling." *American Educational Research Journal 54*: 59–87.

Asparouhov, T., and B. O. Muthén. 2009. "Exploratory structural equation modeling." *Structural Equation Modeling: A Multidisciplinary Journal 16*: 397–438.

Bagozzi, R. P., and T. F. Heatherton. 1994. "A general approach to representing multifaceted personality constructs: Application to state self-esteem." *Structural Equation Modeling: A Multidisciplinary Journal 1*: 35–67.

Bonifay, W., S. P. Lane, and S. P. Reise. 2016. "Three concerns with applying a bifactor model as structure of psychopathology". *Clinical Psychological Science 5*: 184–186.

Broomfield, D., and J. Bligh. 1998. **"**An evaluation of the 'short form' course experience questionnaire with medical students.**"** *Medical Education 32*(4): 367–369.

Byrne, M., and B. Flood. 2003. “Assessing the teaching quality of accounting programmes: An evaluation of the course experience questionnaire.” *Assessment and Evaluation in Higher Education 28*(2): 135–145.

Canivez, G. L. 2016. “Bifactor modeling in construct validation of multifactored tests: Implications for multidimensionality and test interpretation”. In *Principles and methods of test construction: Standards and recent advancements,* edited by K. Schweizer and C. DiStefano, 247–271. Gottingen, Germany: Hogrefe.

Curtis, D. D. 2005. “Comparing classical and contemporary analyses and rasch measurement.” In *Applied Rasch Measurement: A book of exemplars*, edited by S. Alagumali et al. 179–195. The Netherlands: Springer.

Elphinstone, L. 1989. “*The development of the Course Experience Questionnaire.”* M. Ed thesis., The University of Melbourne: Centre for the Study of Higher Education, Melbourne.

Espeland, V., and O. Indrehus. 2003. “Evaluation of students’ satisfaction with nursing education in Norway”. *Journal of Advanced Nursing 42*(3): 226–236.

Franke, G., and M. Sarstedt. 2019. “Heuristics versus statistics in discriminant validity testing: A comparison of four procedures”. *Internet Research 29*(3): 430-447.

Fornell, C., and D. F. Larcker. 1981. “Evaluating structural equation models with unobservable variables and measurement error”. *Journal of Marketing Research 18:* 39–50.

Ginns, P., M. Prosser, and S. Barrie. 2007. “Students’ perceptions of teaching quality in higher education: The perspective of currently enrolled students.” *Studies in Higher Education 32*(5): 603–615.

Henseler, J., C. M. Ringle, and M. Sarstedt. 2015. “A new criterion for assessing discriminant validity in variance-based structural equation modelling”. *Journal of the Academy of Marketing Science 43*(1): 115–135.

Jansen, E., J. van der Meer, and M. Fokkens-Bruinsma. 2013. “Validation and use of the CEQ in the Netherlands”. *Quality Assurance in Education 21*(4): 330–343.

Kreber, C. 2003. “The relationship between students’ course perception and their approaches to studying in undergraduate science courses: A Canadian experience”. *Higher Education Research & Development 22*(1): 57–75.

Marsh, H. W., P. Ginns, A. J. Morin, and B. Nagengast . 2011. “Use of response ratings to benchmark universities: Multilevel modeling of responses to the Australian course experience questionnaire (CEQ).” *Journal of Educational Psychology 103*(3): 733–748.

Marsh, H. W., A. J. Morin, P. D. Parker, and G. Kaur. 2014. “Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis.” *Annual Review of Clinical Psychology* 10: 85–110.

Marsh, H. W., B. Muthén, T. Asparouhov, O. Lüdtke, A. Robitzsch, A. J. S. Morin, and U. Trautwein. 2009. “Exploratory structural equation modeling, integrating CFA and EFA: Application to students’ evaluations of university teaching”. *Structural Equation Modeling 16*(3): 439–476.

McInnis, C., P. Griffin, R. James, and H. Coates. 2001. “*Development of the course experience questionnaire (CEQ)*.” Centre for the Study of Higher Education and Assessment Research Centre. Melbourne: University of Melbourne.

Morin, A. J., A. K. Arens, and H. W. Marsh. 2016. “A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality.” *Structural Equation Modeling: A Multidisciplinary Journal 23*(1): 116–139.

Morin, A. J. S., and T. Asparouhov. 2018. “*Estimation of a hierarchical Exploratory Structural Equation Model (ESEM) using ESEM-within-CFA*.” Montreal, QC: Substantive Methodological Synergy Research Laboratory.

Muthén, B., and D. Kaplan. 1992. “A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model.” *British Journal of Mathematical and Statistical Psychology 45:* 19–30.

Ramsden, P. 1991. “A performance indicator of teaching quality in higher education: The Course Experience Questionnaire.” *Studies in Higher Education 16*: 129–150.

Ramsden, P., and N. J. Entwistle. 1981. “Effects of academic departments on students' approaches to studying”. *British Journal of Educational Psychology 51*(3): 368-383.

Richardson, J. T. E. 1994. “A British evaluation of the course experience questionnaire.” *Studies in Higher Education 19*(1): 59–68.

Richardson, J. T. E. 2009. “What can students’ perceptions of academic quality tell us? Research using the Course Experience Questionnaire.” In *The Routledge international handbook of higher education,* edited by M. Tight, K. H. Mok, J. Huisman, and C. C. Morphew , 199–210. Routledge International Handbooks of Education.

Slade, S., C. Baik, M. Bearman, A. Carbone, M. Hughes-Warrington, D. L. Neumann, and C.D. Smith. 2014. “*Systematic review: What is reported regarding the development of instruments which assess teaching quality in higher education*?”. Australian Government Office for Learning and Teaching.

Spooren, P., F. Vandermoere, R. Vanderstraeten, and K. Pepermans. 2017. “Exploring high impact scholarship in research on student’s evaluation of teaching (SET).” *Educational Research Review 22*: 129–141.

Tabachnick, B. G., and L. S. Fidell. 2001. “*Computer-assisted research design and analysis”*. Boston, MA: Allyn & Bacon.

Watts, A. L., H. E. Poore, and I. D. Waldman. 2019. “Riskier tests of the validity of the bifactor model of psychopathology.” *Clinical Psychological Science 7*(6): 1285–1303.

Waugh, R. F. 1998. “The Course Experience Questionnaire: A Rasch Measurement Model analysis”. *Higher Education Research & Development 17*(1): 45-64.

Wilcox, R. R. 2005. “*Introduction to robust estimation and hypothesis testing*.” Burlington, MA: Elsevier Academic Press.

Wilson, K. L., A. Lizzio, and P. Ramsden. 1997. “The development, validation and application of the Course Experience Questionnaire”. *Studies in Higher Education 22*(1): 33-53.

Table 1. Goodness-of-fit statistics and information criteria for the ICM-CFA and ESEM models estimated on the CEQ23

Models	χ^2	Df	CFI	TLI	RMSEA	90% CI	WRMR
Conventional ICM-CFA							
First-order (ICM-CFA)	887.902	220	.884	.867	.070	.065 - .075	1.525
Higher-order (H-CFA)	887.520	225	.885	.871	.069	.064 - .074	1.58
Bifactor (B-CFA)	628.674	207	.927	.911	.057	.052 - .062	1.263
ESEM							
First-order (ESEM)	388.653	148	.958	.929	.051	.045 - .057	.728
Higher-order (H-EwC)	377.814	153	.961	.936	.049	.043 - .055	.742
Bifactor (B-EwC)	250.362	130	.979	.959	.039	.031 - .046	.546

Table 2. Inter-factor correlations for the CFA and ESEM models

Models	**Inter-factor correlations**										**Median**
	F1-F2	F1-F3	F1-F4	F1-F5	F2-F3	F2-F4	F2-F5	F3-F4	F3-F5	F4-F5	
Conventional CFA											
First-order (ICM-CFA)	.431	.681	.340	.466	.445	.165	.393	.584	.702	.478	.456
Higher-order (H-CFA)	.311	.671	.391	.511	.449	.261	.342	.563	.737	.429	.312
ESEM											
First-order (ESEM)	.291	.336	.198	.332	.214	.059	.272	.435	.496	.358	.439
Higher-order (H-EwC)	.129	.348	.223	.312	.224	.144	.201	.387	.541	.347	.268

Note. F1 = AAS, F2 = AWS, F3= GTS, F4 = GSS, F5=CGS.

Table 3. Parameter estimates for the H-EwC and B-EwC models of the CEQ23

	H-EwC					**B-EwC**					
Items	AAS	AWS	GTS	GSS	CGS	**G**lobal	AAS	AWS	GTS	GSS	CGS
AAS8	.517	.038	.000	.218	-.072	.320	.461	.032	-.012	.164	-.040
AAS12	.624	.114	.284	-.094	-.010	.518	.530	.083	.149	-.125	-.007
AAS19	.473	.065	.021	-.026	.095	.349	.379	.091	-.099	-.069	.038
AWS4	-.011	.813	-.190	.041	.025	.092	-.002	.812	.002	.052	.075
AWS14	-.031	.405	.408	-.020	.066	.593	-.136	.349	-.003	-.101	-.092
AWS20	.201	.657	-.057	-.133	.025	.205	.155	.618	-.024	-.132	.027
AWS23	.046	.700	.017	.021	.008	.279	.072	.637	-.004	-.033	.027
GTS3	.165	-.042	.426	.170	.055	.535	.137	-.062	.257	.126	.046
GTS7	-.076	.002	.523	.069	.035	.488	-.100	-.035	.220	.024	-.027
GTS15	.172	.237	.495	-.046	.058	.639	.088	.176	.145	-.102	-.035
GTS16	.170	-.035	.347	.047	.260	.644	.082	-.083	-.014	-.026	.088
GTS17	.055	-.006	.635	.107	-.046	.542	.047	-.031	.472	.076	-.012
GTS18	.003	.025	.628	.110	.051	.612	-.029	-.015	.390	.082	.026
GSS2	.012	-.050	-.007	.613	.135	.335	.061	-.035	.135	.553	.163
GSS5	-.050	.020	.042	.655	.000	.273	.021	.046	.219	.614	.079
GSS9	.051	-.096	.100	.573	.041	.490	-.003	-.119	-.185	.452	-.077
GSS10	-.074	.043	.120	.743	.008	.434	-.035	.050	.138	.652	.029
GSS11	-.044	-.104	.186	.640	-.137	.401	-.076	-.113	-.078	.520	-.196
GS21	.043	-.018	-.075	.645	.093	.396	-.004	-.051	-.163	.521	.025
CGS1	-.077	.030	-.044	-.018	.637	.344	-.080	.021	.028	-.002	.478
CGS6	-.087	.017	.105	.167	.420	.421	-.089	.003	.068	.142	.301
CGS13	.249	.127	-.177	-.003	.723	.468	.232	.118	-.004	.009	.594
CGS23	-.185	-.024	.224	-.004	.555	.547	-.246	-.059	-.085	-.041	.322
Higher-Or.	.448	.289	.776	.498	.697						

Note. AAS = Appropriate assessment scale; AWS: Appropriate workload scale; GTS: Good teaching scale; GSS: Generic skills scale; CGS: Clear goals and standards scale; H-EwC: single higher-order hierarchical ESEM-within-CFA (EwC); B-EwC: bifactor EwC. The acronyms preceding the number of the item indicate the scale to which it belongs.

Table 4. Convergent and discriminant validity (using Fornell and Larcker's [1981] criteria)

	CR (*ρC)*	**AVE**	**F1-AAS**	**F2-AWS**	**F3-GTS**	**F4-GSS**	**F5-CGS**
Factor 1 - AAS	.561	.301	.549	.017	.121	.050	.097
Factor 2 - AWS	.705	.380	.129	.616	.050	.021	.040
Factor 3 - GTS	.756	.340	.348	.265	.583	.150	.293
Factor 4 - GSS	.799	.397	.223	.177	.402	.630	.120
Factor 5 - CGS	.621	.588	.312	.184	.418	.279	.767
Higher-order f.	.934	.379					

Note: CR: composite reliability; AVE: average variance extracted; Square root of AVE is on the diagonal, estimated latent construct correlations are under the diagonal, and shared variances are above the diagonal.

Table 5. Discriminant validity: Heterotrait-monotrait (HTMT) ratios

HTMT	F1_ASS	F2_AWS	F3_GTS	F4_GSS
F1_AAS				
F2_AWS	.415 CI_{95} [.313 - .518]			
F3_GTS	.178 CI_{95} [.096 - .261]	.453 CI_{95} [.362 - .543]		
F4_GSS	.490 CI_{95} [.389 - .591]	.709 CI_{95} [.610 - .807]	.262 CI_{95} [.191 - .334]	
F5_CGS	.406 CI_{95} [.299 - .513]	.477 CI_{95} [.378 - .577]	.153 CI_{95} [.081 - .225]	.380 CI_{95} [.293 - .468]

Note: The 95% confidence intervals (CI) of each ratio appear below it, in square brackets.

Table 6. Criterion-related validity: correlations among latent factors and learning outcomes

	OSI	**F4 - GSS**
Factor 1 - AAS	.354**	.296**
Factor 2 - AWS	.196**	.166**
Factor 3 - GTS	.472**	.510**
Factor 5 - CGS	.354**	.296**
Factor 4 - GSS	.482**	na
Higher-order f.	.530**	na

Note. OSI: students' overall satisfaction with their course quality; AAS = Appropriate assessment scale; AWS: Appropriate workload scale; GTS: Good teaching scale; GSS: Generic skills scale; CGS: Clear goals and standards scale. ** = p <.001