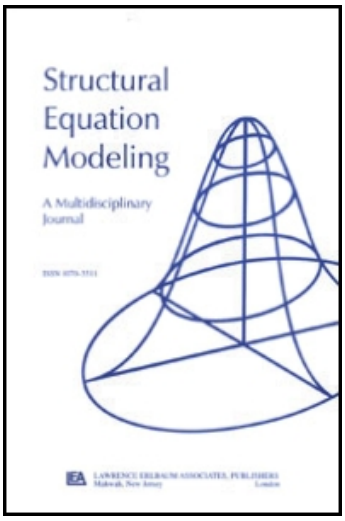

Structural Equation Modeling: A Multidisciplinary Journal
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653699

Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching

Herbert W. Marsh [a]; Bengt Muthén [b]; Tihomir Asparouhov [c]; Oliver Lüdtke [d]; Alexander Robitzsch [e]; Alexandre J. S. Morin [f]; Ulrich Trautwein [d]
[a] Department of Education Studies, Oxford University, [b] University of California, Los Angeles [c] Muthén & Muthén, Los Angeles [d] Max Planck Institute for Human Development, Berlin [e] Institute for Educational Progress, Berlin [f] University of Sherbrooke,

Online Publication Date: 01 July 2009

To cite this Article Marsh, Herbert W., Muthén, Bengt, Asparouhov, Tihomir, Lüdtke, Oliver, Robitzsch, Alexander, Morin, Alexandre J. S. and Trautwein, Ulrich(2009)'Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching',Structural Equation Modeling: A Multidisciplinary Journal,16:3,439 — 476
To link to this Article: DOI: 10.1080/10705510903008220
URL: http://dx.doi.org/10.1080/10705510903008220

Structural Equation Modeling, 16:439–476, 2009

ISSN: 1070-5511 print/1532-8007 online
DOI: 10.1080/10705510903008220

Psychology Press
Taylor & Francis Group

Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching

Herbert W. Marsh
Department of Education Studies, Oxford University

Bengt Muthén
University of California, Los Angeles

Tihomir Asparouhov
Muthén & Muthén, Los Angeles

Oliver Lüdtke
Max Planck Institute for Human Development, Berlin

Alexander Robitzsch
Institute for Educational Progress, Berlin

Alexandre J. S. Morin
University of Sherbrooke

Ulrich Trautwein
Max Planck Institute for Human Development, Berlin

This study is a methodological-substantive synergy, demonstrating the power and flexibility of exploratory structural equation modeling (ESEM) methods that integrate confirmatory and exploratory factor analyses (CFA and EFA), as applied to substantively important questions based on multidimentional students' evaluations of university teaching (SETs). For these data, there is a well established ESEM structure but typical CFA models do not fit the data and substantially inflate

Correspondence should be addressed to Herbert W. Marsh, Department of Educational Studies, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, UK. E-mail: herb.marsh@education.ox.ac.uk

correlations among the nine SET factors (median *r*s = .34 for ESEM, .72 for CFA) in a way that undermines discriminant validity and usefulness as diagnostic feedback. A 13-model taxonomy of ESEM measurement invariance is proposed, showing complete invariance (factor loadings, factor correlations, item uniquenesses, item intercepts, latent means) over multiple groups based on the SETs collected in the first and second halves of a 13-year period. Fully latent ESEM growth models that unconfounded measurement error from communality showed almost no linear or quadratic effects over this 13-year period. Latent *multiple indicators multiple causes* models showed that relations with background variables (workload/difficulty, class size, prior subject interest, expected grades) were small in size and varied systematically for different ESEM SET factors, supporting their discriminant validity and a construct validity interpretation of the relations. A new approach to higher order ESEM was demonstrated, but was not fully appropriate for these data. Based on ESEM methodology, substantively important questions were addressed that could not be appropriately addressed with a traditional CFA approach.

This investigation is a methodological-substantive synergy (Marsh & Hau, 2007). Methodologically, we provide a demonstration of the strength of new exploratory structural equation modeling (ESEM) methods (Asparouhov & Muthén, this issue) that integrate many of the advantages of confirmatory factor analyses (CFA), structural equation modeling (SEM), and exploratory factor analyses (EFA). Substantively we demonstrate the power and flexibility of the ESEM approach as applied to the multiple dimensions of students' evaluations of university teaching (SETs). We begin with an introduction to ESEM, highlighting aspects that are particularly relevant to this investigation (more technical aspects of ESEM are presented in the Methods section and in Asparouhov & Muthén, 2009). We then move to an overview of relevant aspects from the SET literature, with a particular emphasis on a multidimensional perspective. Finally, in the actual analyses that bring together the methodological and substantive themes, we provide new substantive findings that fundamentally rely on ESEM methodology and demonstrate why traditional approaches to CFA would not be appropriate. In this respect, the overarching theme of the article is the importance of substantive-methodological synergy.

Many psychological instruments have an apparently well-defined EFA structure, but cannot be represented adequately within a CFA approach. Typically this is the result of their factor structures not being consistent with the highly restrictive independent clusters model (ICM) typically used in CFA studies in which each item is allowed to load on one factor and all nontarget loadings are constrained to be zero. Although there are many methodological and strategic advantages to ICM-CFAs, they are sometimes inappropriate and many strategies used to compensate for this inappropriateness are dubious, counterproductive, misleading, or simply wrong (see Asparouhov & Muthén, 2009; Browne, 2001). Furthermore, the misspecification of zero factor loadings usually leads to distorted factors with overestimated factor correlations. This can subsequently lead to distortions in structural relations.

DOES ICM-CFA WORK FOR CONVENTIONAL PSYCHOLOGICAL INSTRUMENTS?

In related research, Marsh (2007a; Marsh, Hau, & Grayson, 2005) proposed the following "strawperson" claim:

> Conventional CFA goodness of fit criteria are too restrictive when applied to most multifactor rating instruments. It is my experience that it is almost impossible to get an acceptable fit (e.g., CFI, RNI, TLI > .9; RMSEA < .05) for even "good" multifactor rating instruments when analyses are done at the item level and there are multiple factors (e.g., 5–10), each measured with a reasonable number of items (e.g., at least 5–10 per scale) so that there are at least 50 items overall.

Marsh placed this claim on SEMNET, an electronic mail network devoted to the discussion of SEM issues, and invited the 1,500 members to provide counterexamples. Although a number of interesting points were raised in response to this strawperson claim, no one offered a published counterexample. This suggests that there are many psychological instruments routinely used in applied research that do not even meet minimum criteria of acceptable fit according to current standards. This led Marsh, Hau, et al. (2005; Marsh, 2007a) to question the appropriateness of the new, even more demanding cutoff values (e.g., Hu & Bentler, 1999). An alternative explanation particularly relevant to this investigation is that the typical ICM-CFA structure used to evaluate psychological measures is frequently inappropriate.

Why do researchers persist with apparently inappropriate ICM-CFA models? Because of the recent dominance of CFA approaches to factor analysis, applied researchers have persisted with dubious approaches to CFA in the mistaken believe that EFA approaches were no longer acceptable. These misconceptions have been reinforced by the erroneous beliefs that many of the methodological advances associated with CFAs (e.g., goodness-of-fit assessment, complex error structures, growth modeling, latent mean structures, differential item functioning, tests of the full mean and measurement structures over multiple groups or time, introduction of method factors, bifactor models) are not possible when latent constructs are inferred on the basis of EFAs rather than CFAs. In this investigation we demonstrate how it is possible to apply EFA in a rigorous manner that allows researchers to define more appropriately the underlying factor structure of their constructs than is typically possible in a CFA model, and still rely on the advanced statistical applications typically associated with CFAs and SEM—particularly multiple-group tests of full measurement invariance.

Multiple Group Analysis

The evaluation of model invariance over different groups (e.g., gender or ethnicity) or over different occasions for the same group is widely applied in SEM studies (Jöreskog & Sörbom, 1979; Meredith, 1993; Meredith & Teresi, 2006). Indeed, such tests of invariance might be seen as a fundamental advantage of CFA and SEM over EFA. Although related multiple group methods have been proposed in EFA settings (e.g., Cliff, 1966; Meredith, 1964), they mainly focus on the similarity of factor patterns. However, the ESEM model can be extended to multiple group analyses, where the ESEM model is estimated separately for each group and some parameters can be constrained to invariant across those groups. Here we demonstrate multigroup ESEM tests of invariance that include tests of latent mean differences in EFA factors, item intercepts useful in evaluating differential item functioning, and tests of full measurement invariance of EFA factors, including tests of factorial invariance and models of strong and weak measurement invariance.

Historically, the two main traditions in testing multiple group invariance are based on the invariance of covariance structures and on the invariance of the latent mean structure. Both

approaches typically begin with a model with no invariance of any parameters (configural invariance) and then evaluate the invariance of factor loadings.

From the factor analytic perspective, researchers typically analyze covariance matrices (Marsh, 1994). The initial and primary focus is on the invariance of factor loadings. When there is also an interest in the relations among the factors and on SEM path coefficients, these tests of factorial invariance can be extended to the invariance of the factor variance–covariance matrix. Typically this factor analysis perspective places less emphasis on testing the invariance of item uniquenesses (measurement error), because the focus is on latent traits purged of measurement error. In the traditional factor analysis approach based on covariance structure analyses, information about item means is typically not available. This precludes tests of the invariance of item intercepts and latent means, as well as tests of differential item functioning.

From the measurement invariance perspective, researchers typically begin with a mean augmented covariance matrix. As in the factor analytic perspective, the initial focus is on the invariance of the factor loadings. *Weak measurement invariance* (or pattern invariance) requires that factor loadings be invariant over groups or measurement occasions. An important distinction between the two approaches is the focus on item intercepts associated with differential item functioning. *Strong measurement invariance* requires that the indicator means (i.e., the indicator intercepts) and factor loadings are invariant over groups. In the evaluation of measurement invariance, there is also more emphasis on the item uniquenesses. *Strict measurement invariance* thus requires that, in addition to invariant factor loadings and intercepts, item uniquenesses are also invariant across groups. This implies that group differences on the manifest item-level variables are explained by group differences on the latent factor. As the focus of measurement invariance is typically on evaluation of a single, unidimensional construct, less emphasis is placed on the relations among multiple constructs.[1]

Although measurement invariance and factorial invariance might be seen as distinct, they are clearly very closely related issues (e.g., Meridith & Teresi, 2006). In practice, however, both approaches could be operationalized in relation to a common set of tests defining distinct levels of invariance from the perspective of either CFA or ESEM. For purposes of this investigation, we have operationalized a taxonomy of 13 partially nested models varying from the least restrictive model of configural invariance with no invariance constraints to a model of complete invariance that posits strict invariance as well as the invariance of the latent means and of the factor variance–covariance matrix (Table 1). Although these tests are presented in a logical sequence, the order does not form a strict hierarchy. Although all models except the configural invariance model (Model 1) assume the invariance of factor loadings, it is possible to test, for example, the invariance of indicators' uniquenesses with or without the invariance of the variance–covariance matrix. However, models with freely estimated indicator intercepts and

[1] Although not a focus of this investigation, it is also useful to consider issues of measurement invariance in relation to the terminology from item response theory (see Marsh, 2007a; Marsh & Grayson, 1994). Each measured variable (t) is related to the latent construct (T) by the equation $t = a + b\mathrm{T}$ where b is the slope (or discrimination) parameter that reflects how changes in the observed variable are related to changes in the latent construct and a is the intercept (or difficulty) parameter that reflects the ease or difficulty in getting high manifest scores for a particular measured variable. Unless there is complete or at least partial invariance of both the a and b parameters across the multiple groups, the comparison of mean differences across the groups might be unwarranted. In relation to the taxonomy of models in Table 1, tests of the IRT a parameter are represented by tests of the invariance of factor loadings, and the IRT b parameter is represented by tests of the item intercepts.

TABLE 1
Taxonomy of Multiple Group Tests of Invariance Testable with ESEM

Model	*Parameters Constrained to Be Invariant*
1	None (configural invariance)
2	FL (weak factorial/measurement invariance)
3	FL, Unq
4	FL, FVCV
5	FL, INT (strong factorial/measurement invariance)
6	FL, Unq, FVCV
7	FL, Unq, INT (strict factorial/measurement invariance)
8	FL, FVCV INT
9	FL, UNQ, FVCV INT
10	FL, INT, FMn (latent mean invariance)
11	FL, UNQ, INT, FMn (manifest mean invariance)
12	FL, FVCV, INT, FMn
13	FL, UNQ, FVCV, INT, FMn (complete factorial invariance)

Note. FL = factor loadings; FVCV = factor variance–covariances; INT = item intercepts; Unq = item uniquenesses; FMn = factor means. Models with latent factor means freely estimated constrain intercepts to be invariant across groups, whereas models where intercepts are free imply that mean differences are a function of intercept differences.

freely estimated latent means are not identified, so that the invariance of these parameters cannot be tested in the same model. Indeed, the primary purpose of models with intercepts invariant and means freely estimated is to test whether mean differences at the level of individual items can be explained in terms of differences in latent means.

ESEM MIMIC Models: An Alternative to Multiple Group Analyses

Although the multiple group approach to measurement invariance is reasonable when there is a grouping variable with a small number of discrete groups (e.g., male–female, experimental–control), it is not practical for variables that are continuous or have many categories (e.g., time), nor for studies that simultaneously evaluate many different contrast variables (age, gender, experimental–control) and their interactions. Kaplan (2000; see also Jöreskog & Sörbom, 1988; Marsh, Ellis, Parada, Richards, & Heubeck, 2005; Marsh, Tracey, & Craven, 2005; B. Muthén, 1989) described a MIMIC model, which is like a multivariate regression model in which latent variables are predicted by discrete contrast or grouping variables that are each represented by a single indicator. This approach is clearly stronger than a traditional multivariate analysis of variance (MANOVA) approach that is based on manifest variables, assumed to be measured without error. The MIMIC approach is also much more flexible than the traditional MANOVA approach in allowing a combination of continuous and discrete independent variables and their interactions. Although it is more like a multiple regression approach to analysis of variance (ANOVA; see J. Cohen, Cohen, West, & Aiken, 2003), the MIMIC model has the important advantage that the dependent variables can be latent variables based on multiple indicators corrected for measurement error. Although interaction terms can easily be included in a MIMIC

model if both interacting predictors are single-indicator variables, more complex models are required to model interactions between latent constructs based on multiple indicators (Marsh, Wen, & Hau, 2004, 2006).

Compared to the multiple group invariance tests of measurement invariance, the MIMIC model has some important strategic advantages. Particularly in applied research based on often modest sample sizes for each group, the MIMIC model is much more parsimonious. Also, it allows researchers to consider multiple independent variables and continuous independent variables (without having to recode them to form a small number of discrete groups) that would typically become unmanageable in multiple group analyses. However, there are important limitations as well. Although the MIMIC model provides tests of the invariance of item intercepts (and associated differential item functioning issues), it typically does not include tests of the invariance of factor loadings, uniquenesses, and the factor variance–covariance matrixes like those outlined in Table 1. However, Marsh, Tracey, and Craven (2005) proposed a hybrid approach based on the juxtaposition of multiple-group tests of measurement invariance and the MIMIC model. All models were based on CFA models in which the dependent variables were CFA factors based on multiple indicators. They began with a series of multigroup invariance tests, showing support for strong measurement invariance—with a particular emphasis on the observed variables' intercepts—in separate analyses of each of three different grouping variables. They then constructed MIMIC models in which the dependent variables were regressed on linear and nonlinear components of a continuous variable (age) and interactions between each of the grouping variables. In this way they combined the rigor of the multiple group tests of measurement invariance with the flexibility and parsimony of the MIMIC model. Although based on CFA factors, this hybrid approach is easily pursued in an ESEM approach, juxtaposing results based on the taxonomy of models in Table 1 with a MIMIC model with ESEM factors. Here, as elsewhere in this article, we argue that the ESEM approach might be more appropriate for many applied studies in which the ICM-CFA approach is not able to fit the data adequately.

Fully Latent Growth Models

In his investigation we extend the ESEM model based on this hybrid (MIMIC/measurement invariance) approach to incorporate a MIMIC ESEM latent growth model. Thus, instead of testing for measurement invariance over a small number of discrete occasions in the multiple-group approach, we treat occasions as a continuous time variable in a MIMIC model. Importantly, the starting point for this latent growth model is the individual items. This contrasts with many applications of growth modeling in which the constructs are represented by manifest scale scores (or sets of scores for multidimensional constructs) based on an unweighted average of responses consistent with a typically untested model of the underlying factor structure. We refer to these models as manifest growth models and distinguish them from fully latent growth models.

There are potentially serious problems with manifest growth models. First there is a typically implicit assumption that the a priori model used to compute scale scores fits the data, but this assumption is rarely tested as part of the model. Furthermore, when the construct consists of multiple dimensions, there are many applications in which the ICM-CFA model most closely related to manifest scales scores does not fit the data and can substantially distort the results. In particular, if cross-loadings consistent with an ESEM approach are required to fit the data,

then ICM-CFA latent factors and manifest scores based on this approach are likely to inflate correlations among the factors and undermine the usefulness of a multidimensional perspective. Second, even when there is support for an a priori factor (ICM-CFA) structure, it is unlikely that a simple unweighted average of the multiple indicators provides the optimal representation of the latent construct. Third, even if the first two assumptions are reasonable, the manifest growth model confounds unreliability in the measurement at the level of individual items with stability or instability of the construct over time. Fourth, implicit in the application of this manifest growth model is the Herculean assumption of strict measurement invariance for the construct over the multiple occasions considered in the growth model. This follows in that the comparison of manifest means requires strict measurement invariance whereas the comparison to latent means based on multiple indicators would only require strong measurement invariance (see earlier discussion). Particularly in developmental studies (e.g., academic achievement during school years) in which growth modeling is frequently applied, it is highly unreasonable to assume strict measurement invariance or even to test for it when the tests used on the different occasions are not even the same (i.e., achievement tests used for young children are not the same as those used for older children). Although there is unlikely to be any fully satisfactory solution to these issues, here we present a hybrid ESEM approach that combines advantages of the multiple group invariance tests of measurement invariance and the MIMIC model that is apparently stronger, more rigorous, and more flexible than most approaches.

In summary, although there are still many applications for which pure CFA models are preferable, the new ESEM framework, which even allows for the simultaneous estimations of CFA and EFA models, provides the applied researcher greater flexibility in evaluating the factor structure underlying answers to potentially good instruments that might otherwise be deemed as suspect within pure CFA models.

SUBSTANTIVE APPLICATION: STUDENTS' EVALUATIONS OF TEACHING EFFECTIVENESS

SETs are commonly collected in U.S. and Canadian universities; are increasingly being used in universities throughout the world (e.g., Marsh, 2007c; Marsh & Roche, 1997; Watkins, 1994); are widely endorsed by teachers, students, and administrators; and have stimulated much research spanning nearly a century. Numerous studies have related SETs to a variety of validity criterion measures broadly accepted by classroom teachers (e.g., learning inferred from classroom and standardized tests, student motivation, plans to pursue and apply the subject, positive affect, experimental manipulations of specific components of teaching, ratings by former students, classroom observations by trained external observers, and teacher self-evaluations of their own teaching effectiveness).

Based on extensive reviews of this vast literature consisting of many thousands of research articles, Marsh (1982a, 1983, 1984,1987, 2007b, 2007c; Marsh & Dunkin, 1992, 1997; Marsh & Roche, 1997, 2000) and others (Braskamp, Brandenburg, & Ory, 1985; Cashin, 1988; Centra, 1993; P. A. Cohen, 1980, 1981; Costin, Greenough, & Menges, 1971; de Wolf, 1974; Feldman, 1977, 1978, 1989a, 1989b, 1997; McKeachie, 1997; Remmers, 1963; Richardson, 2005; Rindermann, 1996) concluded that SETs based on well-designed instruments and collected in an appropriate manner tend to be (a) multidimensional; (b) reliable and stable; (c) primarily a

function of the instructor who teaches a course rather than the course that is taught; (d) valid in relation to a variety of indicators of effective teaching; (e) relatively unaffected by a variety of variables hypothesized as potential biases; and (f) seen to be useful by students for use in course selection, by administrators for use in personnel decisions, and by faculty as feedback about teaching. The focus of this investigation is on the multidimensionality of SETs, a characteristic that appears fundamental to their appropriate use.

A Multidimensional Perspective

Researchers and practitioners (e.g., Abrami & d'Apollonia, 1991; Cashin, 1988; Feldman, 1997; Marsh, 2007c; Marsh & Roche, 1993; Renaud & Murray, 2005) agree that teaching is a complex, multidimensional activity comprising multiple interrelated components (e.g., clarity, interaction, organization, enthusiasm, feedback). Hence, it should not be surprising that SETs, like the teaching they are intended to evaluate, should also be multidimensional. This is especially important given the fact that SETs are generally designed as formative or diagnostic feedback tools intended to contribute to the improvement of teaching. As such, they should reflect teaching multidimensionality and target specific aspects needing improvement (e.g., a teacher can be organized but lacking enthusiasm). However, poorly worded, double-barrelled, or inappropriate items will not provide useful information because they will be difficult to interpret, and scores averaged across an ill-defined assortment of items will offer no basis for knowing what is being measured and for targeting specific areas of improvement. Indeed, valid measurement requires a continual interplay among theory, research, and practice and a careful determination of the components that are to be measured.

Marsh and Dunkin (1997) noted three overlapping approaches to the identification, construction, and evaluation of multiple dimensions in SET instruments: (a) empirical approaches such as factor analyses and multitrait-multimethod analyses; (b) logical approaches based on the analysis of the content of effective teaching and of the purposes the ratings are intended to serve, supplemented by reviews of previous research and feedback from students and instructors (see Feldman, 1976); and (c) theoretical approaches based on models of teaching and learning. In practice, most instruments are based on either of the first two approaches—particularly the second. Fortunately, the SET literature contains examples of instruments developed on the basis of more systematic approaches that result in a well-defined EFA factor structure (e.g., Centra, 1993; Jackson et al., 1999; Marsh, 1987, 2007c; Marsh & Dunkin, 1997; Richardson, 2005). Indeed, EFAs of these instruments did demonstrate that they measured distinct components of teaching effectiveness and that there is considerable overlap in the factors measured in each of the various instruments (see Marsh, 1987, 2007c).

The Students' Evaluations of Educational Quality Instrument

Strong support for the multidimensionality of SETs comes from research based on the Students' Evaluations of Educational Quality (SEEQ) instrument (Marsh, 1982a, 1987, 2007c; Marsh & Dunkin, 1997; Marsh & Hocevar, 1991a; Richardson, 2005). The SEEQ measures nine factors (see Appendix A). To develop the SEEQ, a large item pool was first obtained from a literature review, from SET instruments already used, and from interviews with faculty members and students about what they considered to be effective teaching. Students and teachers

were asked to rate the importance of the proposed items; teachers were asked to judge the potential usefulness of the items as a basis for feedback, and students also provided open-ended comments that were examined to determine if important aspects had been excluded. These criteria, along with psychometric properties, were used to select items and revise subsequent versions, thus supporting the content validity of SEEQ responses. Marsh and Dunkin (1992, 1997; Marsh & Roche, 1994) also demonstrated that the content of the SEEQ factors was consistent with general principles of teaching and learning, with a particular emphasis on theory and research in adult education that is most relevant to higher education settings. As noted by Richardson (2005), the SEEQ continues to be the most widely used instrument in published research that provides a strong empirical, conceptual, and theoretical basis for the SEEQ factors.

Factor-analytic support for the SEEQ scales is particularly strong. The EFA structure of SEEQ has been replicated in many published studies, but the most compelling support was provided by Marsh and Hocevar (1991a). Starting with an archive of more than 40,000 sets of class-average ratings, they defined 21 groups of classes that differed in terms of course level (undergraduate or graduate), instructor rank (teaching assistant or regular faculty), and academic discipline. The nine a priori SEEQ factors were identified in each of 21 separate factor analyses. The average correlation between factor scores based on each separate analysis and factor scores based on the total sample was over .99. Although most SEEQ research has focused on student responses to the instrument, the same nine factors were identified in several large-scale studies of teacher self-evaluations of their own teaching using the SEEQ (Marsh, 1983, 1987, 2007c; Marsh, Overall, & Kesler, 1979).

CFAs have largely superseded traditional applications of EFA, and this has created an interesting problem for SET research. This is an important issue, because different practices in the application of EFA and CFA might give the appearance of inconsistent results if not scrutinized carefully (e.g., Toland & De Ayala, 2005). Given the extensive EFA evidence for SEEQ having a clearly defined, replicable structure, why would CFA provide apparently conflicting results? The resolution of this dilemma is that the CFAs are typically based on a highly restrictive ICM structure in which each item is allowed to load on one and only one factor, whereas EFAs allows each item to cross-load on other factors. The exclusion of significant nonzero cross-loadings in CFA might not only result in a poor fit to the data, but could also distort the observed pattern of relations among the factors and the relations between those factors and other constructs. Although there are advantages in having "pure" items that load on a single factor, this is clearly not a requirement of a well-defined, useful factor structure, nor even a requirement of traditional definitions of "simple structure" in which nontarget loadings are ideally small relative to target loadings but not required to be zero (Carroll, 1953; Holzinger & Harman, 1941; Kaiser, 1958; McDonald, 1985; Thurstone, 1947). The extensive EFA results summarized here clearly demonstrate that the SEEQ factor structure is well defined, replicable over a diversity of settings, and stable over time, whereas the ICM-CFA models (e.g., Toland & De Ayala, 2005) do not provide an appropriate representation of the factor structure. In addressing this issue, Marsh (1991a, 1991b) noted that the reason the ICM-CFA structure did not provide an adequate fit to the data was that many items had minor cross-loading on other factors. He randomly divided a large sample of classes into two groups, used empirical techniques to determine a large number of additional (post hoc) parameters (cross-loadings), and then showed that this post hoc solution cross-validated well with the

second sample. However, Browne (2001) argued that such ad hoc strategies might not be the most effective way to identify cross-loading and suggested that EFA approaches should be used instead. Here we demonstrate how ESEM resolves this apparent dilemma in SET research (and in many other applications), integrating the flexibility of an EFA approach with the power of analyses that typically are conducted within a CFA framework.

Potential Biases in Students' Evaluations

The voluminous literature on potential biases in SETs is frequently atheoretical, methodologically flawed, and not based on well-articulated operational definitions of bias, thus continuing to fuel (and to be fueled by) myths about bias (Feldman, 1997; Marsh, 1987, 2007c; Marsh & Dunkin, 1997; Marsh & Roche, 2000). Marsh listed important methodological problems in this research including (a) the inference of causation from correlation; (b) the reliance on inappropriate units of analysis (the class-average is usually appropriate, whereas the individual student rarely is); (c) the neglect of the multivariate nature of SETs; (d) the reliance on inappropriate or logically inconsistent operational definitions of bias; and (e) inappropriate experimental manipulations.

Proper evaluation of SETs' validity, utility, and potential biases (Marsh & Dunkin, 1992; Marsh & Roche, 1997, 2000) demands the rejection of such flawed research, including narrow criterion-related approaches to bias. Instead, as per validity research, it is important to use a broad construct validity approach to the interpretation of bias, which recognizes that (a) effective teaching and SETs designed to measure it are multidimensional; (b) no single criterion measure of effective teaching is sufficient; and (c) theory, measurement, and interpretations of relations with multiple validity criteria and potential biases should be evaluated critically across different contexts and research paradigms. Recognition of the multidimensionality of teaching and of SETs is fundamental to the evaluation of competing interpretations of SET relations with other variables. Although a construct validity approach is now widely accepted to evaluate support for validity, its potential usefulness for the examination of bias issues has generally been ignored. To these concerns we add the problem that most of this research is based on manifest scores based on a typically untested factor structure that fail to control for measurement error—the focus of this investigation.

Marsh and Dunkin (1997; see also Centra, 1993; Marsh, 1987, 2007c) reviewed several large studies of the multivariate relationship between a comprehensive set of background characteristics and multidimensional SETs. In two such studies (see Marsh, 1987), 16 background characteristics explained about 13% of the variance in the set of SEEQ dimensions, but the size and even the direction of these relations varied substantially depending on the SEEQ factor. This research suggested that SETs were positively correlated with higher prior interest in the subject matter, higher expected grades, and higher levels of perceived workload and difficulty, whereas some SEEQ factors were negatively related to class size. Path analyses demonstrated that prior subject interest had the strongest impact on SETs, and that this variable also accounted for about one third of the expected-grade effect. Expected grades also had a negative effect on workload and difficulty because students expecting to receive lower grades perceived the course as more difficult.

Importantly, many researchers—particularly critics of SETs—assume that these correlations necessarily represent bias. In fact, the results are simple correlations that are consistent with

many different interpretations—at least some of which imply that SETs are valid and not biased (see Marsh, 1987, 2007c). Support for a bias hypothesis, as with the study of validity, must be based on a construct validation approach. Indeed, it is ironic that consumers of SET research who have been so appropriately critical of studies claiming to support the validity of SETs have not applied the same level of critical rigor to the interpretation of potential biases in SETs. If a potential biasing factor actually does have a valid influence on teaching effectiveness and this influence is evident in different indicators of teaching effectiveness (e.g., SETs, teacher self-evaluations, student motivation, subsequent course choice, student learning based on objective test scores), then it is possible that the influence reflects support for the validity of SETs (i.e., a valid source of influence in teaching effectiveness is reflected in SETs) rather than a bias. If a background variable has a substantial effect on the specific SET components to which it is most logically related (e.g., class size and rapport with individual students) but has little or no relation to other SET components (e.g., organization) and if this pattern of relations is consistent across multiple methods of measuring teaching effectiveness (e.g., SETs and teacher self-evaluations), again this influence might reflect the validity of SETs rather than a bias. This still leaves the tricky question of how to control for such differences most appropriately when interpreting SETs, but this is a separate question to the most appropriate interpretation of relations between SETs and potentially biasing factors. Thus, for example, apparently no one would argue that student learning as articulated in multisection validity studies (often taken to be the best criterion of validity) is a bias to student ratings rather than a source of validity or that student learning should be partialled from SETs to provide a more valid summary of the SETs.

A full discussion and evaluation of empirical SET research about bias is beyond the scope of this investigation (see Marsh, 2007c). However, it is important to emphasize one important limitation of existing research. Studies based on the SEEQ responses and, apparently, other SET research, are based on a manifest approach in which student ratings are represented by either a single score (a unidimensional perspective) or a set of manifest scores designed to reflect each of the multiple dimensions of teaching effectiveness. Even when there is good support for the a priori factor structure on which the manifest factor scores are constructed, these scores are very limited in comparison with the reliance on latent constructs based on multiple indicators and taking into account measurement error. In addition, the latent variable method based on multiple indicators provides more rigorous tests of the assumptions of measurement invariance implicit in such analyses (see earlier discussion and taxonomy of models in Table 1). Although stronger tests of measurement invariance were traditionally available in a CFA approach, the traditional ICM-CFA approach was typically unable to fit the data in the first place. Due in part to this limitation, SET researchers have typically resorted to suboptimal manifest approaches rather than taking advantage of the important advances in statistical strategies available within the CFA approach. In this respect, ESEM provides a potentially attractive alternative, allowing the use of latent EFA factors in combination with statistical strategies generally reserved to CFAs.

THIS INVESTIGATION: A SUBSTANTIVE METHODOLOGICAL SYNERGY

The substantive orientation of this investigation is to rigorously evaluate the multidimensional perspective of SETs, on the basis of a large archive of class-average responses to the SEEQ

instrument (evaluations of 30,444 classes collected over a 13-year period). We thus demonstrate advanced ESEM statistical analyses in relation to substantively important issues, demonstrating the power and flexibility of the ESEM approach. Importantly, we demonstrate that the ESEM structure fits the data better than competing ICM-CFA models, and that ICM-CFA models systematically distort the latent SEEQ factors in ways that undermine their usefulness and support for a multidimensional perspective. Based on the ESEM solution, we then pursue a rigorous test of measurement invariance of the SEEQ responses based on our new taxonomy of 13 models (Table 1) of multiple group invariance starting with no invariance constraints (configural invariance) to complete invariance (including factor loadings, item intercepts, item uniquenesses, factor variance–covariances, and latent means). We then develop fully latent growth models in which we evaluate systematic changes over 13 years in the ESEM latent factors. Next we apply a MIMIC model to latent ESEM factors to evaluate linear and nonlinear effects of background variables on latent ESEM factors. Finally, we evaluate the higher order structure of SEEQ responses based on a new model developed especially for ESEM applications.

METHODS

The Research Context

Data come from an archive of SETs based on the SEEQ instrument (Marsh, 1982a, 1984, 1987, 2007b, 2007c; Marsh & Bailey, 1993; Marsh & Hocevar, 1991a, 1991b). This archive contains class-average ratings for more than 40,000 classes collected over a 13-year period at one large private, research-oriented university in the United States. For purposes of this investigation, data were 30,444 class-average sets of rating based on responses by at least 10 students, including all undergraduate and graduate-level courses taught by regular faculty. Typically SEEQ instruments were distributed to faculty shortly before the end of each academic term, administered by a student in the class or by administrative staff according to standardized written instructions, and taken to a central office where they were processed. Although an academic unit's participation in this program was voluntary, the university required that all units systematically collect some form of SETs and did not consider any personnel (e.g., tenure, promotion, merit) recommendations that did not include SETs. Thus, most academic units that used SEEQ required all teachers to be evaluated in all courses. Although the SETs at this university have a long history of being broadly accepted, readily available, and widely used, there was no systematic program of teacher development or intervention based on the SETs other than feedback based on SEEQ. Although based on results from a single university, the extensive set of published results based on data from this university (Marsh, 1982a, 1984, 1987; Marsh & Roche, 2000) are broadly consistent with findings from other SET research (Marsh, 1987, 2007b, 2007c; Marsh & Dunkin, 1992, 1997; Marsh & Roche, 1994, 1997).

The ESEM Model

In a basic version of the ESEM model, all parameters can be identified with the maximum likelihood (ML) estimation method. However, when more than one factor is posited ($m > 1.0$),

further constraints are required to achieve an identified solution (Asparouhov & Muthén, this issue). The estimation of the ESEM model consists of several steps. In the first step an SEM model is estimated using the ML estimator. For each block of EFA factors the factor variance–covariance matrix is specified as an identity matrix ($\Psi = I$), giving $m(m + 1)/2$ restrictions. The EFA loading matrix for the block (Λ) has all entries above the main diagonal (i.e., for the first m rows and column in the upper right corner of factor loading matrix, Λ), fixed to 0, providing remaining $m(m - 1)/2$ identifying restrictions. This initial, unrotated model provides starting values that can be subsequently rotated into an EFA model with m factors. The asymptotic distribution of all parameter estimates in this starting value model is also obtained. Then, for each block of EFA factors, the ESEM variance covariance matrix is computed (based only on $\Lambda\Lambda' + \Theta$ and ignoring the remaining part of the model). In M*plus*, multiple random starting values are used in the estimation process to protect against nonconvergence and local minimums in the rotation algorithms. Although a wide variety of orthogonal and oblique rotation procedures are available (e.g., varimax, quartimin, geomin, target, equamax, parsimax, and oblimin), the choice of the most appropriate procedure is to some extent still an open research area.

Multiple Group Analysis

With ESEM models it is possible to constrain the loadings to be equal across two or more sets of EFA blocks in which the different blocks represent multiple discrete groups or multiple occasions for the same group. This is accomplished by first estimating an unrotated solution with all loadings constrained to be equal across the groups. If the starting solutions in the rotation algorithm are the same, and no loading standardizing is used, the optimal rotation matrix will be the same as well as the subsequent rotated solutions. Thus obtaining a model with invariant rotated Λ^* amounts to simply estimating a model with invariant unrotated Λ, a standard task in ML estimation.

For an oblique rotation it is also possible test the invariance of the factor variance–covariance matrix (Ψ) across the groups. To obtain noninvariant Ψs an unrotated solution with $\Psi = I$ is specified in the first group and an unrestricted Ψ is specified in all other groups. Note that this unrestricted specification used here means that Ψ is not a correlation matrix as factor variances are freely estimated. It is not possible in the ESEM framework to estimate a model where in the subsequent groups the Ψ matrix is an unrestricted correlation matrix, because even if in the unrotated solution the variances of the factors are constrained to be 1, in the rotated solution they will not be 1. However, it is possible to estimate an unrestricted Ψ in all but the first group and after the rotation the rotated Ψ can be constrained to be invariant or varying across groups. Similarly, when the rotated and unrotated loadings are invariant across groups, it is possible to test the invariance of the factor intercept and the structural regression coefficients. These coefficients can also be invariant or varying across groups simply by estimating the invariant or group-varying unrotated model. However, in this framework only full invariance can be tested in relation to parameters in Ψ and Λ in that it is not possible to have measurement invariance for one EFA factor but not for the other EFA factors belonging to the same EFA block. Similar restrictions apply to the factor variance covariance, intercepts, and regression coefficients, although it is possible to have partial invariance in the ε matrix of residuals. (It is however, possible to have different blocks of ESEM factors such that invariance constraints are

imposed in one block, but not the other). Furthermore, if the ESEM model contains both EFA factors and CFA factors, then all of the typical strategies for the CFA factors can be pursued with the CFA factors, and these CFA factors can be related to ESEM factors.

Goodness of Fit

In applied CFA and SEM research, there is a predominant focus on indexes that are sample size independent (e.g., Marsh, 2007a; Marsh, Balla, & Hau, 1996; Marsh, Balla, & McDonald, 1988; Marsh, Hau, & Grayson, 2005) such as the root mean squared error of approximation (RMSEA), the Tucker–Lewis Index (TLI), and the comparative fit index (CFI). Thus, for consistency with previous works, these three indexes routinely provided by M*plus* (L. K. Muthén & Muthén, 2008) will be reported, as well as the χ^2 test statistic and an evaluation of parameter estimates. The TLI and CFI vary along a 0-to-1 continuum and values greater than .90 and .95 typically reflect acceptable and excellent fit to the data. RMSEA values of less than .05 and .08 reflect a close fit and a reasonable fit to the data, respectively (Marsh, Hau, & Wen, 2004). Although we rely on these guidelines in this investigation, more research is needed on their appropriateness for ESEM studies for which the number of estimated parameters is typically substantially greater than the typical ICM-CFA study (and degrees of freedom differences in nested models like those considered here can be very large).

Bentler (1990) noted the usefulness of testing a series of nested models. Any two models are nested so long as the set of parameters estimated in the more restrictive model is a subset of the parameters estimated in the less restrictive model. Under appropriate assumptions, the difference in χ^2s between two nested models has a χ^2 distribution and so can be tested in relation to statistical significance. For purposes of model comparison, tests of the relative fit of models testing more or fewer invariance constraints are of greater importance than the absolute level of fit for any one model. However, like the chi-square test statistic itself, some widely used fit indexes like the CFI are monotonic with complexity such that more complex models in a nested sequence always fit better than a less complex model. Thus, for example, Cheung and Rensvold (2001) and Chen (2007) suggested that if the decrease in fit for the more parsimonious model is less than .01 for incremental fit indexes like the CFI, then there is reasonable support for the more parsimonious model. Alternatively, Marsh (2007a) argued that some widely used indexes (e.g., TLI and RMSEA) incorporate a penalty for parsimony so that it is possible for a more parsimonious model to have a better fit value than a less parsimonious model (i.e., the gain in parsimony is greater than the loss in fit). Hence, for these indexes, the more parsimonious model would be supported if the fit index was as good as or better than that for the more complex model. Importantly, however, these are all rough guidelines and should not be interpreted at "golden rules" (Marsh, Hau, & Grayson, 2005; also see Marsh, Hau, Balla, & Grayson, 1998); ultimately there is a degree of subjectivity and professional judgment required in the selection of a "best" model (Marsh et al., 1988). When comparing the relative fit of different models, it is particularly useful to formulate a set of nested or partially nested models specifically designed to evaluate particular aspects of interest like our new taxonomy of models designed to test measurement invariance (see Table 1).

Typically, information criterion indexes like the Akaïke Information Criterion (AIC) are not given much attention in CFA/SEM research because they are so sample-size dependent. However, as noted in the Marsh, Hau, and Grayson (2005) review, sample-size dependency

is not an appropriate criticism of the information criterion indexes in that these indexes are explicitly intended to be highly sample size dependent because more information is available when sample sizes are larger. From a statistical perspective, this is justified because it is appropriate to consider more complex models when the sample size is larger—there is less danger in capitalizing on chance. Thus, Browne (2000) argued that the sample size dependence of information indexes like the Akaïke's Information Criterion (AIC) is appropriate from a cross-validation perspective; to select the model whose parameter estimates are most trustworthy in relation to a particular sample size. The rationale for this family of indexes is that the appropriate level of model complexity and goodness of fit depends on sample size. When the sample size is small, so that sampling error is large, parsimonious models based on relatively fewer estimated parameters will cross-validate better than more complex models based on a relatively larger number of estimated parameters. However, as sample size increases, so that sampling error becomes increasingly small, more complex models will cross-validate more accurately.

Information indexes each have two components: one reflecting approximation discrepancy, which is monotonically related to model complexity, and an opposing estimation penalty reflecting estimation discrepancy (sampling error). Hence, the estimation penalty for each of these indexes is a monotonically decreasing function of sample size and a monotonically increasing function of complexity (the number of estimated parameters in a set of nested models), and has an asymptotic value of zero. Hence, the penalty for parsimony is zero for a sufficiently large sample, but the speed at which each index approaches this value varies for different information indexes. Here we consider three information criterion indexes routinely provided by M*plus* (L. K. Muthén & Muthén, 2008): the AIC, the Bayesian Information Criterion (BIC), and the sample-size adjusted BIC (corBIC). Lower values on these indexes are generally taken to reflect better fit of one model to the data as compared to a model with higher values. However, because of the very large sample size in this investigation ($N = 30{,}444$) and correspondingly small levels of sampling fluctuation, the rationale for these indexes to control for overparameterization in relation to sample size and sampling error has less appeal than it would in studies based on smaller sample sizes. For this reason, we do not place as much emphasis on these indexes for purposes of this investigation, but emphasize that more research is needed to provide appropriate guidance to applied researchers—particularly in the application of the ESEM model.

RESULTS

SEEQ Factor Structure: ESEM versus CFA

We begin with critical analyses used to determine whether the ESEM model does really provide a better fit to the data then a traditional ICM-CFA model. Indeed, unless the ESEM model is better than the ICM-CFA model in terms of goodness of fit and construct validity of the interpretation of the factor structure, then there might be little purpose to pursuing the ESEM approach. Although we argue that the ESEM approach frequently does provide a better fit to the data, it is important to test this assumption in each application. Hence, the starting point of this investigation is the critical comparison between the ESEM and CFA models in terms of

goodness-of-fit indexes (Table 2) and a detailed evaluation of the parameter estimates (Tables 3 and 4).

The ICM-CFA solution does not provide an acceptable fit to the data (e.g., TLI = .871; CFI = .887; RMSEA = .111; see Table 2). Furthermore, there are apparent problems with the parameter estimates. The factor loadings are all substantial and highly significant (Table 3), and the nontarget loadings are necessarily constrained to be zero in the ICM structure. However, many of the factor correlations are so high that they call into question the ability of the instrument to appropriately distinguish between the factors that the SEEQ is intended to measure (Table 3). More specifically, the median factor correlation is .72 (range = .02–.87) and only the correlations involving the workload and difficulty factor are less than .55. Of the remaining correlations, only a few are less than .7 and many are greater than .8. Because an important aim of the SEEQ instrument is to provide diagnostic feedback in relation to particular components of teaching effectiveness and to support the discriminant validity of SETs in relation to other constructs, these large correlations are a serious limitation. When faced with such a poor fitting model, the typical approach is to use modification indexes to free up sufficient parameters to achieve an adequate fit or to pursue potentially dubious techniques to hide the misfit of the model (e.g., forming item parcels masking the misfit, eliminating items that contribute to the misfit even when they are important in terms of content validity). However, even if such techniques were pursued, they would probably not resolve the problem of the inflated correlations.

The ESEM model does fit the data well (e.g., CFI = .961; TLI = .927; RMSEA = .084; see Table 2; also Model TGESEM in Appendix B) and the fit is clearly much better than the corresponding CFA model. Particularly important for purposes of this investigation, the sizes of correlations among factors are substantially smaller for the ESEM solution (median r = .32; range = −.06–.52). Hence the largest ESEM correlation (.52) was smaller than any of the 28 correlations in the CFA model other than those involving the workload and difficulty factor. Particularly this substantial difference in results based on the two models provides a dramatic demonstration of how the ICM-CFA approach can distort the size of relations among the factors by constraining all cross-loading to be zero.

In summary, the ICM-CFA model does not fit these data adequately and the ESEM approach does. In this respect the results provide the initial and most important test for the appropriateness of the ESEM model. Having established this, we now demonstrate the versatility and flexibility of the ESEM model, and how we can address substantively important issues that could not otherwise be appropriately addressed.

Robustness and Stability of SEEQ Factor Structure and Latent Means

How stable is the SEEQ factor structure over time? Do mean ratings increase over time—a grade inflation effect? Because the same instrument was used continuously over 13 years, there is the concern as to whether the nature of the constructs remained constant over time. Whereas the first question focuses primarily on the invariance of the factor loadings, the second question deals with full measurement invariance to compare the means appropriately. To pursue these substantively important questions and demonstrate the power of the ESEM model, we applied the taxonomy of 13 models to these data (Table 1; see also the M*plus* inputs for MGI1 and MGI13 in Appendix B). For these analyses, we divided the set of evaluations into two groups

TABLE 2
Summary of Goodness of Fit Statistics for All Models

Model	*Chi-Square/df*	*NFParm*	*CFI*	*TLI*	*RMSEA*	*AIC*	*BIC*	*corBIC*	*Description*
Total group (TG) models									
TGCFA	198367.199/524	141	.887	.871	.111	105,526	106,699	106,251	Total group CFA
TGESEM	68211.182/316	349	.961	.927	.084	−24,213	−21,308	−22,418	Total group ESEM
Multiple (two) group invariance (MGI; also see Table 1)									
MGI1	68806.046/632	698	.961	.927	.084	−28,019	−22,209	−24,427	IN = none (FMn = 0)
MGI2	70384.682/866	464	.960	.945	.073	−26,908	−23,046	−24,521	IN = FL (FMn = 0)
MGI3	71374.339/901	429	.960	.947	.072	−25,988	−22,418	−23,781	IN = FL, Uniq (FMn = 0)
MGI4	70950.837/911	419	.960	.948	.071	−26,432	−22,944	−24,276	IN = FL, FVCV (FMn = 0)
MGI5	71453.950/892	438	.960	.946	.072	−25,891	−22,245	−23,637	IN = FL, INT(FMns free)
MGI6	71966.657/946	384	.959	.949	.070	−25,486	−22,290	−23,510	IN = FL, Uniq, FVCV, (FMns free)
MGI7	72449.486/927	403	.959	.948	.071	−24,965	−21,611	−22,892	IN = FL, Uniq, Inter (FMns free)
MGI8	72020.305/937	393	.959	.948	.071	−25,414	−22,143	−23,392	IN = FL, FVCV, Inter (FMns free)
MGI9	73042.135/972	358	.959	.950	.070	−24,463	−21,483	−22,621	IN = FL, FVCV, Inter, Uniq (FMns free)
MGI10	71720.571/901	429	.960	.947	.072	−25,642	−22,071	−23,435	IN = FL, INT, FMn
MGI11	72715.466/936	394	.959	.948	.071	−24,717	−21,438	−22,690	IN = FL, Uniq, Inter, FMn
MGI12	72288.232/946	384	.959	.949	.070	−25,165	−21,968	−23,189	IN = FL, FVCV, INT, FMns
MGI13	73309.433/981	349	.959	.950	.070	−24,213	−21,308	−22,418	IN = FL, FVCV, INT, Uniq, FMn
MIMIC growth (MIMICGRW) models									
MIMICGRW 1A	70070.625/394	376	.960	.929	.076	225,691	228,821	227,626	Lin, Quad, Cub on 9 factors
MIMICGRW 1B	67886.443/316	463	.961	.914	.084	223,681	227,535	226,063	Lin, Quad, Cub on 35 items
MIMICGRW 1C	70606.676/421	349	.960	.933	.074	226,173	229,078	227,969	Lin, Quad, Cub all = 0
MIMIC background (MIMIC-BCK) effect models									
Linear									
MIMIC-BCK1A	73935.064/394	385	.958	.926	.078	215,043	218,247	217,024	PInt, Enrl, EGrd on 9 factors
MIMIC-BCK 1B	65807.572/316	463	.963	.918	.083	207,071	210,925	209,454	PInt, Enrl, EGrd on 35 items
Linear and quadratic									
MIMIC-BCK2A	76166.611/472	403	.957	.927	.073	634,414	637,769	636,488	PInt, Enrl, EGrd (L&Q) on 9 factors
MIMIC-BCK 2B	65012.643/316	559	.964	.907	.082	623,573	628,226	626,449	PInt, Enrl, EGrd (L&Q) on 35 items
MIMIC-BCK 2C	97180.168/526	349	.946	.917	.078	655,320	658,225	657,116	PInt, Enrl, EGrd (L&Q) = 0
Higher order (HO) factor models									
HO1A	80627.969/323	342	.954	.915	.090	−11,811	−8,964	−10,051	1 HO factor

Note. CFI = comparative fit index; TLI = Tucker–Lewis Index; NFParm = number of free parameters; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; corBIC = sample-size adjusted BIC; RMSEA = root mean squared error of approximation; SRMR = standardized root mean square residual. For multiple group invariance models, IN = means the sets of paramaters constrained to be invariant across the multiple groups: FL = factor loadings; FVCV = factor variance–covariances; INT = item intercepts; Uniq = item uniquenesses; FMn = factor means. Background variables are: Pint = prior subject interest; Enrl = enrollment; EGrd = class-average expected grade.

TABLE 3
ESEM Solution: Nine ESEM Factors Based on Responses to 35 SEEQ Items

Factors	*Factor Loadings*								
Items	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Learning/value (F1)									
Q1	.47	.15	.02	.10	.01	.30	.14	.10	.16
Q2	.65	.06	.03	.09	.05	.11	.07	.07	.16
Q3	.68	.07	.05	.02	.05	−.00	.15	.07	.13
Q4	.47	.04	.16	.19	.04	−.25	.04	.11	.14
Enthusiasm (F2)									
Q5	.08	.59	.07	.12	.19	.02	.11	.05	.08
Q6	.06	.72	.06	.09	.08	.03	.09	.07	.08
Q7	.10	.63	.11	−.02	.07	−.05	.12	.11	.06
Q8	.15	.54	.13	.19	.03	.01	.07	.12	.07
Group interaction (F3)									
Q13	.05	.11	.79	.02	.06	.00	.03	.06	.08
Q14	.06	.05	.81	−.01	.10	−.03	.09	.04	.07
Q15	.07	.10	.57	.18	.15	.00	.10	.10	.08
Q16	.06	.06	.67	.04	.18	−.02	.11	.09	.06
Organization/clarity (F4)									
Q9	.13	.15	.16	.53	.06	−.08	.09	.10	.10
Q10	.05	.09	.04	.67	.07	.00	.09	.11	.12
Q11	.12	.02	.04	.53	.07	.03	.06	.19	.20
Q12	.03	.04	−.13	.55	.04	.03	.23	.14	.06
Individual rapport (F5)									
Q17	.07	.14	.18	−.00	.62	−.05	.05	.13	.06
Q18	.04	.06	.08	.06	.77	−.00	.04	.10	.07
Q19	.08	.13	.14	.02	.64	−.00	.05	.14	.06
Q20	−.04	−.01	.01	.13	.65	.04	.12	.11	.13
Workload/difficulty (F6)									
Q32	.01	.03	−.07	.06	−.02	.88	.07	.01	.04
Q33	.07	−.04	.06	.01	.00	.89	−.01	−.03	.05
Q34	−.08	.11	−.10	.04	−.01	.72	.01	.04	.06
Q35	.11	−.04	.02	−.10	.03	.81	−.02	.00	.04
Breadth of coverage (F7)									
Q21	.04	.06	.05	.10	.07	.07	.68	.07	.12
Q22	.08	.10	.02	.12	.04	.00	.67	.06	.12
Q23	.05	.06	.13	.08	.11	−.01	.62	.10	.11
Q24	.25	.10	.09	.03	.03	−.02	.54	.08	.03
Exams/grading (F8)									
Q25	.03	.04	.04	.12	.10	.06	.03	.67	.09
Q26	.04	.04	.04	.01	.11	−.03	.08	.75	.11
Q27	.06	.04	.02	.12	.06	−.03	.06	.65	.15
Assignments/reading (F9)									
Q28	−.03	−.01	.01	−.00	.01	−.01	.05	.00	.94
Q29	.13	.01	.03	.04	.03	.07	−.01	.14	.73
Overall course									
Q30	.41	.20	.04	.18	.04	.07	.08	.17	.18
Overall teacher									
Q31	.16	.36	.08	.27	.13	.04	.10	.15	.08

	Factor Correlations								
	F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*	*F8*	*F9*
F1	1.00								
F2	.46	1.00							
F3	.34	.39	1.00						
F4	.44	.45	.24	1.00					
F5	.26	.41	.46	.33	1.00				
F6	.13	.07	−.09	.03	.00	1.00			
F7	.45	.43	.33	.47	.36	.07	1.00		
F8	.42	.41	.32	.53	.49	.03	.41	1.00	
F9	.50	.34	.29	.46	.35	.16	.42	.52	1.00

Note. ESEM = exploratory structural equation modeling; SEEQ = Students' Evaluation of Educational Quality instrument. The ESEM model was an exploratory factor analysis with 9 SEEQ factors (see model TGESMEM in Model 2 for goodness-of-fit statistics). All parameter estimates are completely standardized. N = 30,444 class-average sets of rating to the 35 SEEQ items. A priori target loadings are shaded in gray.

TABLE 4
ICM-CFA Solution: Nine ESEM Factors Based on Responses to 35 Items From the SEEQ Instrument

Factors	*Factor Loadings*								
Items	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Learning/value									
Q1	.924	.000	.000	.000	.000	.000	.000	.000	.000
Q2	.940	.000	.000	.000	.000	.000	.000	.000	.000
Q3	.931	.000	.000	.000	.000	.000	.000	.000	.000
Q4	.804	.000	.000	.000	.000	.000	.000	.000	.000
Enthusiasm									
Q5	.000	.945	.000	.000	.000	.000	.000	.000	.000
Q6	.000	.964	.000	.000	.000	.000	.000	.000	.000
Q7	.000	.899	.000	.000	.000	.000	.000	.000	.000
Q8	.000	.960	.000	.000	.000	.000	.000	.000	.000
Organization/clarity									
Q9	.000	.000	.935	.000	.000	.000	.000	.000	.000
Q10	.000	.000	.969	.000	.000	.000	.000	.000	.000
Q11	.000	.000	.934	.000	.000	.000	.000	.000	.000
Q12	.000	.000	.802	.000	.000	.000	.000	.000	.000
Group interaction									
Q13	.000	.000	.000	.947	.000	.000	.000	.000	.000
Q14	.000	.000	.000	.963	.000	.000	.000	.000	.000
Q15	.000	.000	.000	.951	.000	.000	.000	.000	.000
Q16	.000	.000	.000	.980	.000	.000	.000	.000	.000
Individual rapport									
Q17	.000	.000	.000	.000	.947	.000	.000	.000	.000
Q18	.000	.000	.000	.000	.966	.000	.000	.000	.000
Q19	.000	.000	.000	.000	.974	.000	.000	.000	.000
Q20	.000	.000	.000	.000	.845	.000	.000	.000	.000
Breadth of coverage									
Q21	.000	.000	.000	.000	.000	.940	.000	.000	.000
Q22	.000	.000	.000	.000	.000	.950	.000	.000	.000
Q23	.000	.000	.000	.000	.000	.951	.000	.000	.000
Q24	.000	.000	.000	.000	.000	.867	.000	.000	.000
Exams/grading									
Q25	.000	.000	.000	.000	.000	.000	.915	.000	.000
Q26	.000	.000	.000	.000	.000	.000	.963	.000	.000
Q27	.000	.000	.000	.000	.000	.000	.937	.000	.000
Assignments/reading									
Q28	.000	.000	.000	.000	.000	.000	.000	.862	.000
Q29	.000	.000	.000	.000	.000	.000	.000	.999	.000
Workload/difficulty									
Q32	.000	.000	.000	.000	.000	.000	.000	.000	.844
Q33	.000	.000	.000	.000	.000	.000	.000	.000	.954
Q34	.000	.000	.000	.000	.000	.000	.000	.000	.702
Q35	.000	.000	.000	.000	.000	.000	.000	.000	.862
Overall course									
Q30	.960	.000	.000	.000	.000	.000	.000	.000	.000
Overall teacher									
Q31	.000	.947	.000	.000	.000	.000	.000	.000	.000
	Factor Correlations								
	F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*	*F8*	*F9*
F1	1.000								
F2	.868	1.000							
F3	.855	.854	1.000						
F4	.699	.763	.673	1.000					
F5	.697	.794	.721	.809	1.000				
F6	.836	.821	.831	.715	.723	1.000			
F7	.785	.774	.833	.675	.788	.752	1.000		
F8	.787	.652	.721	.560	.596	.662	.733	1.000	
F9	.282	.148	.119	.021	.065	.160	.095	.279	1.000

Note. ICM-CFA = independent clusters model-confirmatory factor analysis; ESEM = exploratory structural equation modeling; SEEQ = Students' Evaluation of Educational Quality instrument. The CFA model (see model TGCFA in Table 2 for goodness-of-fit statistics) assumed an independent cluster structure in which each of the 35 SEEQ items was allowed to load on only a single latent factor and all nontarget loadings were constrained to be zero (i.e., cross-loadings were not allowed). All parameter estimates are completely standardized. To conserve space, the factor loadings are presented in condensed format such that only the target loading relating each item to its a priori factor is presented (as all nontarget loadings are zero). N = 30,444 class-average sets of rating to the 35 SEEQ items.

based on ratings from the first 7 years (Group 1) and the last 6 years (Group 2)—but later consider an alternative approach in which time (the 13 years) is considered as a continuous variable.

The two-group model with no invariance constraints (MGI1 in Table 2) provides a good fit to the data (e.g., CFI = .961, TLI = .927). Indeed, these fit statistics are approximately the same as those based on the total group ESEM model (TGESEM in Table 2) with twice the degrees of freedom (632 vs. 316) and twice the number of free parameters (698 vs. 349). These results support the configural invariance of the SEEQ: that the same ESEM model is able to fit data from Groups 1 and 2 when no additional invariance constraints are imposed.

Model MGI2 (Table 2) constrains factor loadings to be invariant across the two groups. It is, perhaps, the most important model from both factor analysis and measurement invariance perspectives. Because the number of freely estimated factor loadings in the ESEM model is so substantial, MG2 is much more parsimonious than MGI1: The number of freely estimated parameters drops from 698 to 464. Nevertheless, MGI2 provides a very good fit to the data. Indeed, for the traditional fit indexes that control for parsimony, the fit of MGI2 is systematically better than the fit of MGI1 (e.g., TLI = .945 vs. .927; RMSEA = .073 vs. .084). Although the chi-square value for MGI1 is significantly smaller than that of MG2 (due to the very large *N*), the chi-square/*df* ratio is substantially smaller. Interpretations based on the information criteria are not as clear. The AIC (Table 2) is clearly better for the MGI2 than MGI1, but MGI1 is slightly better based on the corrected BIC and particularly the uncorrected BIC. This apparently reflects the combination in this investigation of the extremely large sample size (so that there is almost no capitalization on chance, and that is the primary focus of information indexes) and the very substantial difference in parsimony between MGI1 and MGI2. The CFI is monotonic with parsimony for nested models, but even here the change in CFIs (.961 vs .960) is considerably less than the .01 value recommended to support interpretations of invariance (Cheung & Rensvold, 2001). In summary, the juxtaposition of Models MGI1 and MGI2 provide clear support for the invariance of the factor loadings, sometimes referred to as weak measurement invariance.

Strong measurement invariance requires that item intercepts—as well as factor loadings—are invariant over groups. The critical comparison is thus between Models MGI2 and MGI5. The change in *df* = 26 represents the 35 new constraints on item intercepts less the 9 latent factor means that are now freely estimated. This model tests whether the 35 intercepts can be explained in terms of 9 latent means. A lack of support for this model would suggest differential item functioning, meaning that differences between items' mean levels in the two groups cannot be solely explained in terms of differences at the latent factor mean levels. Based on traditional fit indexes, the fit of MGI5 is almost equivalent—slightly better than—to the fit of MGI2 (e.g., TLI = .946 vs. .945; RMSEA = .072 vs. .073). However, each of the information criteria is now better (lower) for MGI5 than MGI2, whereas the value of the CFI is approximately the same (.960). In summary, there is good support for strong measurement invariance, suggesting that latent means can appropriately be compared in Groups 1 and 2.

Strict measurement invariance requires that item uniquenesses, item intercepts, and factor loadings are all invariant over the groups. Here, the critical comparison is between models MGI5 and MGI7. The change in *df* = 35 represents the 35 new constraints added on item uniquenesses. A lack of support for this model would suggest that measurement error differs in the two groups. Typically, this would reflect either systematically higher or lower measurement

error in the different groups, but might also reflect a differentiated pattern at the level of the latent construct or individual items. Based on traditional fit indexes that control parsimony, the fit of MGI7 is marginally better than the fit of MGI5 (e.g., TLI = .948 vs. .946; RMSEA = .071 vs. 073). Also, all three of the information criteria are better (lower) for MGI7 than MGI5, whereas the value of the CFI is nearly as good (.959 vs. .960). In summary, there is good support for strict measurement invariance, suggesting that the manifest means can appropriately be compared in Groups 1 and 2.

Invariance of the factor variance–covariance matrix is typically not a focus in studies of measurement invariance because they typically focus on tests of unidimensionality based on a single construct. However, this is frequently an important focus of studies of the invariance of covariance structures—particularly studies of the discriminant validity of multidimensional constructs that might subsequently be extended to include relations with other constructs. Here, the most basic comparison is between Models MGI2 (factor loadings invariant) and MGI4 (factor loadings and factor variance–covariances invariant). The change in $df = 45$ represents the 36 factor covariances and 9 factor variances. Based on fit indexes that control for parsimony, the fit of MGI4 is marginally better than MGI2 (TLI = .948 vs. .945; RMSEA = .071 vs. .073). Also, all three the information criteria are better (lower) for MGI4 than MGI2, whereas the value of the CFI is approximately the same (.960). In summary, there is good support for the invariance of the factor variance–covariance matrix across the two groups.

Tests of the invariance of the latent factor variance–covariance matrix, as is the case with other comparisons, could be based on any pair models in Table 2 that differ only in relation to the factor variance–covariance matrix being free or not, for example: (a) MG1 (FL) versus MG4 (FL, FVCV); (b) MGI3 (FL, Uniq) versus MGI6 (FL, Uniq, FVCV); (c) MGI5 (FL, Inter) versus MGI8 (FL, Inter, FVCV); (d) MGI7 (FL, Inter,Uniq) versus MGI9 (FL, Inter, Uniq, FVCV); (e) GI10 (FL, Inter, FMns) versus MGI9 (FL, Inter, FMNs, FVCV). Although each of these pairs of models differ by $df = 46$ corresponding to the parameters in the variance–covariance matrix, they are not equivalent; support for the invariance of the variance–covariance matrix could be found in some of those comparisons, but not in others. Although we suggest that the comparison between models MGI4 and MGI2 represents the most basic of those comparisons, valuable information can also be obtained from the other comparisons as well. Particularly if there are systematic, substantively important differences in the interpretations based on these comparisons, further scrutiny would be warranted. This is especially true when sample sizes are small because true differences in the factor variance–covariance matrix might be "absorbed" into differences in other parameters that are not constrained to invariance. In this respect, pairs of models that constrain more parameters to be invariant (but differ in terms of constraining or not constraining the factor variance–covariance matrix to invariance) might provide more conservative tests of the invariance hypothesis. Fortunately, this complication is not evident in this investigation as support for the invariance of factor variance–covariance matrix is consistent across each of these alternative comparisons.

Finally, we are now in a position to address the issue of the invariance of the factor means across the two groups. Again, there are several models that could be used to make this comparison: (a) MGI5 (FL, Inter; strong measurement invariance) versus MGI10 (FL, Inter, FMn); (b) MGI7 (FL, Inter, Uniq; strict measurement invariance) versus MGI11 (FL, Inter, Uniq, FMn); (c) MGI8 (FL, Inter, FVCV) versus MGI12 (FL, Inter, FVCV, FMn); (d) MGI9 (FL, FVCV, Uniq, Inter) versus MGI13 (FL, FVCV, Uniq, Inter, FMn). Following the

measurement invariance tradition, we emphasize comparisons with strong and strict invariance assumptions, but suggest that each of these pairs of models warrants consideration, as was the case for multiple models of factor variance–covariance invariance. Indeed, the final model in this set (MGI9 vs. MGI13) apparently provides the most rigorous, conservative test of the assumption of equal means. In this study, it is important to note that all of those comparisons support the invariance of factor means over the two groups, with fit indexes that control for parsimony and the information criteria all showing that the more constrained models fit the data as well or better than the less constrained models. In summary, there is good support for the invariance of factor means over the two groups.

The focus of discussion has been on a microperspective, based on particular pairs of nested models designed to assess the invariance of a particular set of parameters. However, it is also useful to evaluate results from the entire set of models from a more global perspective. Although comparisons based on nonnested models should be interpreted cautiously, the fit indexes emphasized here do not require models to be nested. Particularly for indexes that include a control for parsimony, there is a consistent pattern of results. MGI13, clearly the most restrictive model, has more degrees of freedom (requires less parameter estimates) but still fits the data better than any other models according to indexes that include a control for parsimony (TLI, RMSEA), and all three of the information indexes. Across the 13 models there is very little variation in CFIs (.959–.961, a range of .002) and clearly less than the minimum difference indicative of noninvariance (.01). Particularly for the TLI and RMSEA, there is a dramatic improvement in going from MGI1 to MGI2, and much smaller improvements based on subsequent, more restrictive models. However, for the ESEM model considered here, there is also a massive change in parsimony going from MGI1 to MGI2 ($\Delta df = 234$) compared to that for any other comparisons ($\Delta df = 125$ going from MGI2 to MGI13).

It is also important to note that the extensive taxonomy of models considered here is not exhaustive, and that other more specific models might be useful to consider in particular applications. However, even when more specific models are considered, it is useful to evaluate them in relation to results from this taxonomy with particular emphasis on multiple sets of nested models that facilitate interpretations.

Returning to the substantive focus of this investigation, the results provide clear affirmative answers to each of our research questions. First, there is clear evidence that the SEEQ factor structure is reasonably invariant across the students' evaluations of teaching collected in the first and second half of the 13-year period considered here. Conversely, there is no evidence that the nature of the constructs has changed over this period. In offering this interpretation, our primary focus was on support for MGI2—the test of the invariance of the factor loadings. However, the finding that the factor variance–covariance matrix is also invariant across the two groups also is relevant and provides additional support to this conclusion.

Results of this investigation also support the conclusion there has been no systematic increase or decrease in the mean ratings based on SEEQ over this 13-year period. This finding is substantively important in relation to policy practice. In particular, it is reasonable to compare results based on ratings collected early in the history of SEEQ with those collected subsequently. Although our basis for this conclusion is very strong in relation to a measurement invariance perspective and our taxonomy of 13 models (Table 1), researchers should always be cautious about translating research findings into policy practice. For example, the comparisons considered here are largely cross-sectional in that the classes (teachers and particularly the

students evaluating them) are different in Groups 1 and 2. However, Marsh (2007b) also evaluated the mean stability of ratings for a group of 195 teachers who were evaluated continuously over this 13-year period of time (an average of more than 30 classes per teacher). For this longitudinal comparison (in relation to the teachers), mean differences in the ratings of the same teacher over time were also remarkably stable; 84% of the teachers showed no increase or decrease in ratings, a few showed increases, and a few showed decreases. Teachers who received good ratings early in this period tended to have good ratings across the entire period, whereas those who received poor ratings were consistently rated as poor. In another study, Marsh and Overall (1979) considered ratings by the same set of students who all took the same core courses over a number of different cohorts. In this study, individual students were identified and ratings for each course were collected once at the end of the class and once again at least 2 years after graduation from the program (i.e., the study was longitudinal in relation to the students). Based on ratings for 100 different classes, test-retest correlations were close to the reliabilities of the ratings and there were no systematic differences in the mean level of ratings. The juxtaposition of results from these different perspectives all support the stability of ratings over time.

Mean Stability From a Growth Model Perspective

The results of the multigroup invariance tests provide strong support for the invariance of the SEEQ factor structure over time and a demonstration of the power of ESEM. However, although some grouping variables are truly dichotomous, much potentially valuable information is lost when a small number of groups are formed from a potentially much larger number of groups (or a continuous variable for which the number of groups is theoretically infinite). In particular, in this investigation there are 13 year groups that were divided into two groups for the purposes of the multiple-group tests of invariance. Although in theory it would have been possible to consider all 13 groups in the multigroup analysis, the size of the model would have been unmanageable. Nevertheless, it is possible that systematic differences were masked by this crude classification. To explore this possibility, an ESEM latent growth model was also estimated.

In this ESEM latent growth model we begin with individual items—multiple indicators of each latent construct (MIMIC-GRW1A in Table 2; see also the M*plus* input in Appendix B). As described earlier, this latent growth model has important advantages over traditional manifest growth models that are typically used in applied research. The starting point is the comparison of the total group models already presented (TGCFA and TGESEM in Table 2). These preliminary models demonstrate that the ESEM model provides a good fit to the data and the CFA approach does not.

Next a MIMIC (latent growth) model was tested in which the effects of the 13 year groups were represented by orthogonal contrast variables representing the linear, quadratic, and cubic effects of year. Because factor loadings were essentially the same as in Table 1, we focused specifically on the path coefficients reflecting the growth components associated with the multiple SEEQ factors. Results (Table 5; growth components) demonstrate that the direction, pattern, and effect sizes based on the linear growth component closely parallel those based on the dichotomous year variable in the multigroup tests of measurement invariance considered earlier.

TABLE 5
Changes in ESEM Factors Over 13 Years: Comparison of ESEM Latent Mean Differences Based on Multigroup Tests of Measurement Invariance Treating Years as a Dichotomous Grouping Variable and Polynomial Latent Growth Functions Based on a MIMIC Treating Years as a Continuous Variable

	Models: Latent Mean Differences				*Polynomial Growth Components*		
Factor	*MGI7*	*MGI5*	*MGI8*	*MGI9*	*Linear*	*Quad*	*Cubic*
Learning/value	0.023	0.023	0.021	0.021	0.005	−0.052**	−0.004
Enthusiasm	0.055**	0.055**	0.055**	0.055**	0.027**	−0.029**	−0.002
Organization/clarity	0.060**	0.060**	0.060**	0.060**	0.033**	0.016*	0.010
Group interaction	−0.056**	−0.056**	−0.057**	−0.057**	−0.026**	−0.005	−0.007
Individual rapport	−0.094**	−0.094**	−0.096**	−0.096**	−0.054**	−0.001	−0.004
Workload/difficulty	0.049**	0.049**	0.049**	0.049**	0.034**	−0.025**	0.000
Breadth of coverage	−0.067**	−0.067**	−0.068**	−0.068**	−0.038**	−0.007	0.005
Exams/grading	−0.028*	−0.028*	−0.028*	−0.028*	−0.010	−0.028	0.003
Assignments/reading	0.039**	0.039**	0.039**	0.039**	0.028**	−0.001	0.019**

Note. ESEM = exploratory structural equation modeling. For models of latent mean differences, year was divided into two groups (Group 1 = first half of the 13 years, Group 2 = second half of the 13 years). In these multiple-group tests of measurement variance, latent means were fixed to be zero in Group 1 and freely estimated in Group 2 so that latent means in group represent latent mean differences in the two groups (completely standardized). Latent means were estimated in relation to four alternative models (MGI7, MGI5, MGI8, MGI9; see Table 2). For polynomial latent growth components, the 13 years were treated as a continuous time variable, represented by linear, quadratic, and cubic polynomial growth components. The effects of these growth components on latent ESEM factors were estimated with a MIMIC model (MIMICGRW 1A in Table 2). Latent mean differences based on the multigroup tests are in the metric of Cohen's *d* effect size, whereas the corresponding Cohen's *d* for the linear growth component can be approximated by the translation of correlation into Cohen's *d* by the equation ($d = 2r/\sqrt{(1 - r^2)}$).

$^*p < .05$; $^{**}p$, .001.

Marsh, Tracey, and Craven (2005; see also Grayson, Mackinnon, Jorm, Creasey, & Broe, 2000; Kaplan, 2000) extended the MIMIC model to allow paths from contrast variables to latent variable indicators as well as to the latent factors. As in the multiple-group approach, this allows researchers to test whether there are group differences in individual items beyond what can be explained by differences in overall means of the latent constructs. Kaplan (2000) demonstrated how latent mean differences with the MIMIC model are analogous to those with the traditional multiple-group approach under appropriate conditions. In this application, we posited a model (MIMIC-GRW2 in Table 2) in which the three polynomial growth components led to the 35 indicators (SEEQ items) directly rather than indirectly through their relation to the latent means. This model is less parsimonious in that the 27 parameters (9 SEEQ factors × 3 growth components) are replaced by 105 (35 SEEQ items × 3 growth components) parameters (i.e., $\Delta df = 105 - 27 = 78$). For fit indexes that take into account parsimony, the more parsimonious MIMIC-GRW1 model (TLI = .929, RMSEA = .076) fit the data better than the MIMIC-GRW2 model (TLI = .914, RMSEA = .084) and the difference in CFI (.960 vs. .961) was smaller than the change of .01 typically used to support the less parsimonious model. Interestingly, however, all three information indexes supported the less parsimonious MIMIC-GRW2 in which the growth components affected individual items rather than latent means. Again, this result based in information indexes apparently reflects a combination of

the very large sample size that is idiosyncratic to this study and the substantial difference in parsimony.

Finally, we considered a third model in which all the polynomial growth components (for both latent factors and item intercepts) were constrained to be zero. This model is clearly more parsimonious than either of the other latent growth models. Reflecting this increased parsimony, the results indicated that the fit was somewhat better in terms of fit indexes that take into account parsimony (see the MIMIC-GRW3 model in Table 2: TLI = .933 vs. .927; RMSEA = .074 vs. .076). However, the information indexes led to the selection of the less parsimonious models over the model with growth components constrained to be zero (Table 2).

The MIMIC approach has some important advantages as well as some limitations in relation to the traditional multiple-group approach. In its favor is its flexibility in accommodating reasonably continuous variables that cannot be easily incorporated into the multiple-group tests of measurement invariance—particularly in studies based on modest sample sizes. Here, for example, we represented the 13 year groups with polynomial contrasts. In other applications, it would be possible to use any of the wide variety of contrasts (orthogonal or nonorthogonal)—as well as interactions among main effect contrasts—like those used in traditional ANOVA models. However, there are potentially problematic assumptions that are not easily tested within the MIMIC framework. Thus, for example, there is an implicit assumption that the factor loadings, variance–covariance matrix, and item uniquenesses are invariant across groups, but this cannot be readily tested within the MIMIC framework. The hybrid approach demonstrated here provides a good compromise between the multigroup and MIMIC approaches to rigorous tests of measurement invariance. We began by conducting traditional multiple group analyses in which we evaluated the invariance of various combinations of factor loadings, factor variances, factor covariances, measured variable uniquenesses, and measured variable intercepts. For these multiple-group comparisons, year was "dichotomized" into two groups. We then conducted a MIMIC analysis that included the effects of year treated as a continuous variable. In this hybrid strategy, we were able to pursue rigorous tests of invariance possible with the multiple-group approach, while still taking advantage of the flexibility of the MIMIC approach. Importantly, both approaches provided similar conclusions in this study.

A MIMIC Approach to Effects of Background Variables

Relations with background variables: Potential biases. A critical issue in relation to SETs is the relation between SETs and background and demographic variables considered as potential biases. Among the most frequently considered potential biases are class size (negative bias), prior subject interest (positive bias), expected grades and grading leniency (positive bias), and workload and difficulty (negative bias). There is a growing body of research suggesting that potential biases may reflect a valid influence that is appropriately reflected in the SETs (Marsh, 1987, 2007c; Marsh & Roche, 2000). Critical to this research is the demonstration that background variables have differentiated patterns of relations with well-differentiated multiple dimensions of SETs. However, most research has been based on manifest scores to represent the multiple SET factors, potentially inflating substantially the sizes of correlations among the multiple SET factors and undermining support for discriminant validity (see earlier discussion). Results presented here are apparently the first test of the discriminant validity of multiple SET factors in relation to background characteristics in which the multiple factors are represented as

latent constructs rather than manifest scores. Two issues are particularly relevant in this ESEM application: (a) Is the direction, size, and nature of the relations between a background variable and SETs relatively consistent across different SEEQ factors, suggesting a generalized bias that is independent of content? (b) Are there nonlinear as well as linear effects relating background variables to the SEEQ factors that might complicate or undermine simple interpretations of bias?

Description of ESEM models. In this investigation, we evaluated two sets of ESEM models relating three background variables (prior subject interest, class-average expected grades, and enrollment) to the nine ESEM SEEQ factors. In one set of models we considered the linear effects of the background variables and in a second set of models we included linear and quadratic effects of the background variables (see Tables 2 and 6). For purposes of these analyses, we tested a MIMIC model in which background variables are considered as independent variables and the nine SEEQ factors are dependent variables. Importantly, this is apparently the first such analysis in which the SEEQ factors are represented as latent factors based on multiple indicators in a single model. Because of the inability to model the SEEQ factors within an ICM-CFA framework, previous research has been based on manifest SEEQ scores in models that did not include the multiple indicators of each SEEQ factor. The results reported here thus provide a potentially important substantive contribution as well as a demonstration of the usefulness and flexibility of the ESEM approach.

Earlier we noted that the MIMIC approach is capable of testing some aspects of measurement invariance and might be particularly useful when the "grouping" variable can take on many values or is continuous. Here, the three background variables are all reasonably continuous (i.e., can take on hundreds of values). Because measurement invariance and issues such as differential item functioning typically focus on grouping variables that can take on a small number of discrete values, the issue of measurement invariance in relation to a continuous variable has received less attention. Nevertheless, the concern here is the same as when the grouping variable takes on discrete values. In terms of differential item functioning, the issue

TABLE 6
Relations Between Nine SEEQ Factors and Four Background Variables

	Corr With	*Path Model (Linear Effects)*				*Path Model (Linear and Quadratic Effects)*						
SEEQ Factors	*Wrk/Diff*	*PSI-L*	*CAGE-L*	*ENR-L*	*MultR²*	*PSI-L*	*PSI-Q*	*CAGE-L*	*CAGE-Q*	*ENR-L*	*ENR-Q*	*MultR²*
Learning/value	.121**	.451**	.192**	−.047**	.278**	.442**	−.007	.193**	−.093**	−.046**	.033**	.288**
Enthusiasm	.072**	.082**	.113**	.020**	.022**	.076**	.020**	.112**	−.059**	.022**	.036**	.027**
Grp interact	−.080**	.090**	.189**	−.244**	.125**	.077**	.003	.183**	−.081**	−.245**	.121**	.147**
Ogan/clarity	.027**	.013*	.029**	.030**	.002**	.010	.003	.025**	−.019**	.032**	.013**	.002**
Ind. rapport	.012**	−.017**	.167**	−.182**	.068**	−.027**	−.009	.163**	−.074**	−.183**	.092**	.082**
Breadth cover	.079**	.195**	.142**	.036**	.055**	.160**	−.049**	.143**	−.089**	.034**	.023**	.066**
Exam/grades	.054**	.018**	.217**	−.109**	.067**	.007	.027**	.215**	−.112**	−.107**	.082**	.088**
Assign/read	.158**	.198**	.096**	−.022**	.057**	.194**	−.012	.092**	−.012	−.023**	.037**	.058**
Work/diff	1.000**	.195**	−.251**	−.042**	.086**	.198**	.071**	−.256**	.115**	−.039**	.038**	.108**

Note. Relations among four background variables (Wrk/Diff = workload/difficulty; PSI = prior subject interest; CAGE = class-average grade expectations; ENR = enrollment) and nine student evaluation factors (SEEQ = Students' Evaluation of Educational Quality instrument). As Wrk/Diff is considered a SEEQ factor, correlations between it and the other factors are presented. An exploratory structural equation modeling path model was used to estimate the simultaneous effects of the other three background variable on the set of SEEQ factors. Separate models were conducted on linear (L) effects and the combination of linear (L) and quadratic (Q) effects.

*$p < .05$; **p, .001.

is whether relations between the background variables and each of the 35 SEEQ items can be explained reasonably in terms of the nine SEEQ factors.

In MIMIC-BK1A and MIMIC-BK1B (Table 2; see also the M*plus* input in Appendix B), we compare a more parsimonious model in which paths lead from linear background variables to nine SEEQ factors (1A) with a less parsimonious model with paths from background variables to 35 SEEQ items (1B). The models differ substantially in terms of parsimony in representing these relations (3 background variables × 9 factors vs. 3 background variables × 35 items, $\Delta df = 78$). Evaluation of the goodness-of-fit statistics provides reasonable support for measurement invariance in relation to item intercepts. In particular, indexes that control for parsimony are better for the more parsimonious MIMIC-BK1A (TLI = .927, RMSEA = .073) than MIMIC-BK1B (TLI = .918, RMSEA = .083), as are the small difference in CFIs (i.e., 964 vs. .957, less than .01). However, the information indexes—like the chi-square statistic—favor the less parsimonious MIMIC-BK1B, apparently reflecting the very large sample size. In MIMIC-BK1C, we also evaluated a model in which the effects of all background characteristics were constrained to be zero. Although Model 1C clearly had a worse fit than Model 1A (indicating that there are effects of the background variables), its fit was still surprisingly good (suggesting that the background effects were not large). The second set of models (MIMIC-BK2A–2C in Table 2) is essentially the same, except the set of three background variables was expanded to include linear and quadratic effects of each background variable. However, in terms of the fit indexes, the same pattern of results is evident in those models (Table 2). We now turn to an evaluation of the effects of each of the background variables.

Workload and difficulty relations. We begin with an evaluation of the workload/difficulty factor that, for purposes here, is considered as both a SEEQ factor and as a potential background factor. For this reason, we limit this analysis to the consideration of the correlations observed among the ESEM factors (see Table 3). These results show that there are mostly small, positive relations between the workload/difficulty factor and the other SEEQ factors. However, because of the extremely large sample size, all of these relations are statistically significant ($p < .01$). Only two correlations are greater than .10 ($r = .158$ with assignments/readings, and $r = .121$ with learning/value). The only negative relation is small and involves the group interaction factor ($r = -.08$). These results suggest that teachers teaching more difficult classes and requiring more work from their students are evaluated as more effective in relation to most of the SEEQ factors. Consistent with a multidimensional perspective, these positive effects are largest for the ratings obtained on the learning/value factor, on the readings/assignments factor, and, to a lesser extent, on the breadth of coverage factor. Because of the direction of these relations—positive rather than negative—a bias hypothesis (i.e., that teachers are "rewarded" for making classes easier and requiring less work) is not viable. This pattern of results is also evident in teacher self-evaluations of their own teaching (Marsh, 2007c; Marsh & Roche, 2000), thus supporting the construct validity of the interpretation of workload and difficulty as having a positive influence on SETs.

Enrollment. The linear effects of enrollment (class size) are quite varied, ranging from small positive relations with the breadth of coverage factor, organization/clarity, and enthusiasm to higher negative relations with individual rapport (−.182) and group interaction (−.244). This highly differentiated pattern of relations is consistent with a construct validity interpretation of

the results from a multidimensional perspective. It is, of course, logical that class size would have particularly negative effects on the quality of group interaction and on the individual relationships that students can build with teachers. Similarly, it not surprising that enrollment has small positive relations with the breadth of coverage, organization/clarity, and enthusiasm factors. In each case, similar patterns of relations were also evident in teacher self-evaluations of their own teaching (e.g., Marsh, 1982b, 1987, 2007c).

It is also important to note that enrollment appears to exert consistent positive quadratic effects such that there is a decline in ratings when enrollment rises from small to moderately large classes, but this relationship changes direction when enrollment further increases such that very large classes tend to receive ratings that are as positive as those of small classes. Marsh (1987, 2007c) speculated that for increasingly large classes, particularly effective teachers were chosen to teach these classes who increasingly adopt teaching strategies especially suited to large classes and that their reputation as effective teachers might also attract larger enrollments. This pattern of results is consistent with a construct validity interpretation that enrollment has a negative effect on some aspects of teaching effectiveness (group interactions and individual rapport) that is appropriately reflected in the SEEQ ratings, but that these negative effects might be offset by using more appropriate large-class teaching strategies.

Prior subject interest (PSI). The linear effects of PSI are quite varied, ranging from close to zero to .451. Consistent with a construct validity interpretation, PSI is most substantially related to SEEQ factors to which it is most logically related: particularly learning/value (.451) and, to a lesser extent, assignments/readings (.198), breadth of coverage (.195), and workload/difficulty (.195). Relations with other SEEQ factors are all smaller (less than .10). Again, in support of the construct validity interpretations, relations with teacher self-evaluations show a similar pattern of results (Marsh, 1987, 2007c). Although there is some nonlinearity in the patterns of relations (see Table 6) the quadratic effects are mostly small and many are not even statistically significant. The pattern of relations, and the consistency of these relations based on students' evaluations and teachers' self-evaluations of their own teaching suggest that prior subject interest does have an effect on some aspects of effective teaching (particularly learning and value) that are appropriately reflected in the SEEQ responses.

Class-average expected grades (CAGEs). Interpretations based on CAGEs are particularly complicated in that CAGEs reflect a combination of student learning (better grades reflect better achievement), teacher grading leniency (higher grades reflect more lenient teachers), and the effects of prior characteristics (more able, motivated students get better grades). Positive relations between CAGEs and student ratings reflects the construct validity of SEEQ responses according to the first interpretation, an apparent bias according to the second interpretation, and an incidental, spurious effect according to the third interpretation. Although relations between CAGEs and SEEQ factors are clearly important, a full discussion of the interpretation of these relations is beyond the scope of this article (see Marsh, 2007c). The relations between CAGEs and the SEEQ factors are quite differentiated, varying from moderately negative (workload/difficulty, −.251) to moderately positive (exams/grading, .217; learning/value, .192). Furthermore, there are substantial nonlinear effects such that the CAGEs are positively related to SETs for low to moderately high levels of CAGEs, but that this function then plateaus and declines for very high CAGEs. The pattern of relations—particularly the positive relations

with the learning/value and exams/grading factors and the negative relation with the workload/difficulty factor—is also evident in teacher self-evaluations (Marsh, 1987, 2007c). These nonlinear effects seem inconsistent with a grading-leniency interpretation in that the relation is negative—not positive—for the most "lenient" segment of the function. Marsh (2007c; Marsh & Roche, 2000) also noted that a large body of studies based on the multisection validity paradigm show that class-average objective measures of achievement (based on a common metric that eliminates any grading leniency effect) and SETs tend to be correlated .5 and higher—values even higher than relations between CAGEs and SETs. Nevertheless, there is some evidence that controlling for student ability and prior subject interest does reduce the size of relations between CAGEs and the SEEQ factors. A cautious interpretation of these results suggests that the relations between CAGEs and SEEQ factors reflect student learning (consistent with a validity interpretation) and, perhaps, the effects of prior characteristics (a spurious effect), but that that there is limited support for a grading-leniency bias interpretation.

Summary of ESEM MIMIC background analyses. The analyses conducted in this section demonstrate the strength and flexibility of the ESEM approach in relation to MIMIC models. In particular, analyses along the lines of those presented here have been previously based on manifest scale scores and could not be properly conducted with CFA models based on multiple indicators. These analyses raised new issues such as tests of measurement invariance in relation to continuous variables that can be addressed—at least in part—from an ESEM perspective. Even more clearly than in previous studies, the results indicate that the direction, size, and nature of the effects of background variables vary substantially for different SEEQ factors, and that some of the relations have a substantial nonlinear component. A detailed evaluation of the results of each background variable supported a construct validity interpretation suggesting that each variable had a small effect on the components of teaching effectiveness most logically related to it and, apparently, that these effects were appropriately reflected in SEEQ responses. In summary, these results demonstrate the usefulness of an ESEM approach for methodological-substantive synergy.

Higher Order Factor Analysis

In general it is possible to do a second-order factor analysis based on the EFA model, but this is not readily available in the current version of ESEM—as it is easily accomplished with a CFA approach. In this investigation we explored some alternatives to this traditional approach that could be conducted with ESEM. It would be possible to use a two-stage approach in which the correlations among the first-order factors (e.g., those in Table 3) were the input for a second, entirely separate model. Although potentially useful from a purely descriptive perspective, such two-stage analyses are dubious in terms of model comparisons like those in traditional CFA approaches to higher order factors.

Coupled with this concern is an interesting idiosyncratic feature of the SEEQ instrument. In addition to the SEEQ items designed to measure specific factors, there are two overall rating items—overall rating of the teacher and the course. Although substantially correlated with each other and logically related to many of the different SEEQ factors, these items load on different factors (see Table 3). The overall teacher rating is most strongly related to the teacher enthusiasm factor (.36), but also has statistically significant cross-loadings on each of the other

SEEQ factors. The overall course rating is most strongly related to the learning/value factor (.41), but it also has significant cross-loadings on the other SEEQ factors.

Taking advantage of these overall rating items, we posited a separate model in which the two overall rating items were used to define a separate global teacher effectiveness CFA factor. The remaining 33 SEEQ items were used to define the nine ESEM factors (as before, but excluding the two overall rating items). The CFA global teacher effectiveness factor was posited to influence each of the nine ESEM factors (i.e., arrows go from the global factor to specific factors, as in traditional CFA approaches to higher order factor analysis). Importantly, this higher order factor model (HO1A in Table 2; see also the M*plus* input in Appendix B) is nested under the traditional ESEM factor model (TGESEM in Table 2). In particular, the 18 factor loadings relating the two overall rating items to the nine factors in TGESEM are represented by nine correlations (relating the one global factor and the nine specific factors) and the two factor loadings in HO1A ($\Delta df = 7$). Even though the difference in parsimony between these two models is small, all of the goodness-of-fit indexes suggest that this new global model (HO1A) provides a systematically poorer fit to the data than the first-order ESEM model (TGESEM). Although perhaps disappointing from the perspective of supporting this new approach to higher order factor analysis within an ESEM context, the results are consistent with the multidimensional perspective that is the overarching substantive focus of this investigation.

SUMMARY, DISCUSSION, AND IMPLICATIONS

The substantive orientation of this investigation was to rigorously evaluate the multidimensional perspective to SETs based on SEEQ responses. In accomplishing this, we demonstrated the usefulness and greater flexibility of the ESEM approach in relation to substantively important issues that could not be appropriately evaluated with traditional ICM-CFAs. Based on a very large database (evaluations of 30,444 classes collected over a 13-year period) we showed that the ESEM structure fits the data better than the ICM-CFA approach. Of particular relevance, the ICM-CFA systematically distorted the size of correlations among the latent SEEQ factors. Thus the median correlation among ESEM factors was only .34, whereas that among CFA factors based on the same data was .72. Although such a huge difference might seem surprising, it is consistent with the logic of the ESEM approach. In particular, when a large number of relatively small cross-loadings are constrained to be zero as in the ICM-CFA solution (Table 4), the only way that these cross-loadings can be represented is by inflating the size of correlations. In relation to this investigation—and to SET research more generally—this is a particularly serious problem because it undermines support for (a) the multidimensional perspective that is the overarching rationale for this study, (b) the discriminant validity of the multiple SEEQ factors, and (c) the usefulness of the ratings in terms of providing diagnostic feedback to improve teaching effectiveness. Although dramatically demonstrated in a way that is substantively important in this investigation, we suggest that similar phenomena are likely to occur in most applications where the ICM-CFA model is inappropriate.

In pursuing the methodological aims of this investigation, we demonstrated the flexibility of the ESEM approach and its applicability to a wide variety of different situations that are likely to be useful for applied researchers. In the application of the ESEM approach to tests of multigroup invariance, we developed a new taxonomy of 13 models that more fully integrated

the tests from both the traditional factor-analytic approach and the measurement invariance approach. The ability to test such a detailed taxonomy with ESEM demonstrates the flexibility of the ESEM approach.

In evaluating the question of whether mean SEEQ ratings had changed over the 13 years considered here, we extend the ESEM approach to a truly latent growth model of change. This model offers potentially important advantages over traditional growth models based on manifest scores. Implicit in the application of growth models is the very strong assumption of measurement invariance over time. Because the traditional growth models are based on manifest means, it requires the particularly stringent assumption of strict measurement invariance (i.e., invariance of factor loadings, item intercepts, and item uniquenesses). In contrast, because the latent growth model is based on latent means, it only requires the less stringent assumption of strong invariance (i.e., invariance of factor loadings and item intercepts). However, many applications of growth models in applied research completely ignore all assumptions of measurement invariance, substantially undermining the credibility of interpretations of the results. Indeed, for studies in which completely different measures are used from one occasion to the next, it is difficult to see how it is even possible to conduct rigorous tests of measurement invariance. Particularly in applied research, we suggest that the hybrid approach demonstrated here is a good compromise between the rigor of full multigroup tests of measurement invariance and the flexibility of the MIMIC approach. Again, application of this latent growth model as part of the ESEM approach demonstrates the versatility and flexibility of the ESEM framework.

We also evaluated the relations among the nine ESEM latent SEEQ factors and the linear and quadratic components of background variables posited to represent potential biases to the SEEQ factors. Again, it was important in these models to evaluate the latent SEEQ factors inferred on the basis of multiple indicators. For example, in these MIMIC models we evaluated differential item functioning, demonstrating that effects of the background variables at the level of individual items could be reasonably represented by background effects at the level of latent factors. Although we did not pursue a full application of the hybrid approach to evaluate measurement invariance in relation to each of the background variables considered separately, we note that assumptions of measurement invariance implicit in growth models apply here as well. Thus, the strategies to address these issues used in this investigation should have broad applicability for applied researchers wishing to more fully justify the interpretations of their results from a measurement perspective.

Finally, we proposed an ESEM approach to testing higher order ESEM factor structures. This was developed to overcome the limitations of the ESEM approach in relation to traditional CFA approaches to higher order factor analysis. Although the new higher order factor model did not work particularly well in this application, it did illustrate the flexibility of the ESEM approach. In particular, within a single model, we combined a CFA factor based on one set of items and ESEM factors based on another set of items, and related the two sets of latent constructs. Although we refer to the global factor as a CFA factor, there is really no difference between a CFA factor and an ESEM factor when all indicators (in this case the two overall rating items) load on a single latent factor (global teaching effectiveness). With more items and more than one factor, it would be possible to posit alternative models in which the second set of items could be represented by either ESEM or CFA factors, and relate these factors to the ESEM factors based on the first set of items. Although not pursued in this investigation (but see Muthén & Muthén, 2008), it would also be possible to define two sets of factors—one

based on ESEM and one based on CFA—in relation to responses to the same items. This type of model might be useful for a study in which CFA factors were posited to reflect method effects (e.g., positively and negatively worded items, items from different instruments) whereas trait factors based on the same items were represented as ESEM factors.

Importantly, the analytical strategies demonstrated here could also be applied in traditional ICM-CFA studies. In this respect we present the ESEM model as a viable alternative to the ICM-CFA model, but we do not argue that the ESEM approach should replace the corresponding CFA approach. Indeed, when the basic ICM-CFA model is able to fit the data as well as the ESEM model, there are important strategic advantages to the use of the CFA approach. However, at least in this investigation, the ICM-CFA model was not able to fit the data but the ESEM model was able to do so. In this situation, we suggest that advanced statistical strategies such as multigroup and MIMIC tests of measurement invariance and latent growth models are more appropriately conducted in an ESEM approach than with a traditional ICM-CFA approach. From this perspective, the results of this investigation provide dramatic evidence that an ESEM approach is more appropriate than a traditional ICM-CFA approach for responses to SEEQ. However, based on Marsh's (2007a; Marsh, Hau, & Grayson, 2005) suggestion that apparently almost no multidimensional psychological instruments widely used in practice provide an acceptable fit in relation to an a priori ICM-CFA structure, we suspect that issues identified with SEEQ responses may have broad generalizability. Clearly there is need for further research based on the application of the ESEM approach to other multidimensional constructs.

REFERENCES

Abrami, P. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness: Generalizability of $N = 1$ research. Comment on Marsh (1991). *Journal of Educational Psychology, 30,* 221–227.

Asparohov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*(3), 397–438.

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

Braskamp, L. A., Brandenburg, D. C., & Ory, J. C. (1985). *Evaluating teaching effectiveness: A practical guide.* Beverly Hills, CA: Sage.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44,* 108–132.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111–150.

Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika, 18,* 23–38.

Cashin, W. E. (1988). *Student ratings of teaching: A summary of research.* (IDEA Paper No. 20). Manhattan, KS: Kansas State University, Division of Continuing Education. (ERIC Document Reproduction Service No. ED 302 567)

Centra, J. A. (1993). *Reflective faculty evaluation.* San Francisco, CA: Jossey-Bass.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14,* 464–504.

Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods, 4,* 236–264.

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika, 31,* 33–42.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis. *Research in Higher Education, 13,* 321–341.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51,* 281–309.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity and usefulness. *Review of Educational Research, 41,* 511–536.

de Wolf, W. A. (1974). *Student ratings of instruction in post secondary institutions: A comprehensive annotated bibliography of research reported since 1968* (Vol. 1). Seattle: University of Washington Educational Assessment Center.

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

Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses. *Research in Higher Education, 6,* 223–274.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers and courses: What we know and what we don't. *Research in Higher Education, 9,* 199–242.

Feldman, K. A. (1989a). Association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30,* 583–645.

Feldman, K. A. (1989b). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30,* 137–194.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368–395). New York: Agathon.

Grayson, D. A., Mackinnon, A., Jorm, A. F., Creasey, H., & Broe, G. A. (2000). Item bias in the Center for Epidemiologic Studies Depression scale: Effects of physical disorders and disability in an elderly community sample. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 5,* 273–282.

Holzinger, K. J., & Harman, H. H. (1941). *Factor analysis.* Chicago: University of Chicago Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Jackson, D. L., Teal, C. R., Raines, S. J., Nansel, T. R., Force, R. C., & Burdsal, C. A. (1999). The dimensions of students' perceptions of teaching effectiveness. *Educational and Psychological Measurement, 59,* 580–596.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models.* Cambridge, MA: Abt Books.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7—A guide to the program and applications* (2nd ed.). Chicago: SPSS.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 187–200.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Newbury Park, CA: Sage.

Marsh, H. W. (1982a). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52,* 77–95.

Marsh, H. W. (1982b). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74,* 264–279.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75,* 150–166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76,* 707–754.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11,* 253–388.

Marsh, H. W. (1991a). A multidimensional perspective on students' evaluations of teaching effectiveness: A reply to Abrami and d'Apollonia (1991). *Journal of Educational Psychology, 83,* 416–421.

Marsh, H. W. (1991b). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology, 83,* 285–296.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1,* 5–34.

Marsh, H. W. (2007a). Application of confirmatory factor analysis and structural equation modeling in sport/exercise psychology. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of on sport psychology* (3rd ed., pp. 774–798). New York: Wiley.

Marsh, H. W. (2007b). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99,* 775–790.

Marsh, H. W. (2007c). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Ed.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–384). New York: Springer.

Marsh, H. W., & Bailey, M. (1993). Multidimensionality of students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education, 64,* 1–18.

Marsh, H. W., Balla, J. R., & Hau, K.-T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315–353). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. Smart (Ed.), *Higher education: Handbook on theory and research* (Vol. 8, pp. 143–234). New York: Agathon.

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241–320). New York: Agathon.

Marsh, H. W., Ellis, L., Parada, L., Richards, G., & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment, 17,* 81–102.

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling, 1,* 317–359.

Marsh, H. W., & Hau, K.-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology, 32,* 151–171.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McCardle (Eds.), *Psychometrics: A festschrift to Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341.

Marsh, H. W., & Hocevar, D. (1991a). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7,* 9–18.

Marsh, H. W., & Hocevar, D. (1991b). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7,* 303–314.

Marsh, H. W., & Overall, J. U. (1979). Long-term stability of students' evaluations. *Research in Higher Education, 10,* 139–147.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology, 71,* 149–160.

Marsh, H. W., & Roche, L. A. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30,* 217–251.

Marsh, H. W., & Roche, L. A. (1994). *The use of students' evaluations of university teaching to improve teaching effectiveness.* Canberra, Australia: Australian Department of Employment, Education, and Training.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52,* 1187–1197.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workloads on students' evaluations of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology, 92,* 202–228.

Marsh, H. W., Tracey, D. K., & Craven, R. G. (2005). *Multidimensional self-concept structure for preadolescents with mild intellectual disabilities: A hybrid multigroup-MIMIC approach to factorial invariance and latent mean differences.* Sydney, Australia: SELF Research Centre, University of Western Sydney.

Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9,* 275–300.

Marsh, H. W., Wen, Z., & Hau, K.-T. (2006). Structural equation models of latent interaction and quadratic effects. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 225–265). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52,* 1218–1225.

Meredith, W. (1964). Rotation to achieve factorial invariance. *Psychometrika, 29,* 187–206.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance, *Psychometrika, 58,* 525–543.

Meredith, W., & Teresi, J. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(Suppl. 3), S69–S77.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585.

Muthén, L. K., & Muthén, B. (2008). *M*plus *user's guide.* Los Angeles: Muthén & Muthén.

Remmers, H. H. (1963). Rating methods in research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 329–378). Chicago: Rand McNally.

Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education, 46,* 929–953.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education, 30,* 387–415.

Rindermann, H. (1996). On the quality of students' evaluations of university teaching: An answer to evaluation critique. *Zeitschrift Für Padagogische Psychologie, 10*(3–4), 129–145.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago.

Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65,* 272–296.

Watkins, D. (1994). Student evaluations of teaching effectiveness: A cross-cultural perspective. *Research in Higher Education, 35,* 251–266.

APPENDIX A: 9 SEEQ FACTORS AND THE TWO OVERALL SUMMARY RATINGS

Learning/value	Valuable learning experience, was intellectually stimulating or challenging
Instructor enthusiasm	Instructor displayed enthusiasm, energy, humor, and ability to hold interest
Organization/clarity	Organization and clarity of explanations, materials, objectives, lectures
Group interaction	Students encouraged to discuss, participate, share ideas, and ask questions
Individual rapport	Lecturer accessible, friendly, and interested in students
Breadth	Presentation of broad background, concepts, and alternative approaches and theories
Exam/graded materials	Student perceptions of value and fairness of exams and graded materials
Readings/assignments	Value of assignments in adding appreciation and understanding to subject
Workload/difficulty	Subject difficulty, workload, pace, and hours outside of class
Overall class rating	Overall rating of the course
Overall teacher rating	Overall rating of the teacher

APPENDIX B: SAMPLE M*plus* INPUT FILES FOR SELECTED ESEM MODELS

Total Group ESEM (TGESEM in Table 2; also see Table 3)

```
TITLE: Syntax for Model TGESEMMG1 (see Table 2)
DATA: FILE IS USCSET2YRGRPN=30444NV=52.DAT; !user-specified data file;
```

```
VARIABLE: NAMES ARE q1-q38 q40 year enroll rank per395
   flrn fenth forgn fgrp find fbrd fexam fasgn fwork;
   USEVARIABLES q1-q29 q30 q31 q32-q35; ! variables actually used in the analysis
ANALYSIS:
     type = efa 9 9; !specifies a 9-factor EFA solution;
     ROTATION=GEOMIN(OBLIQUE, .5);
  OUTPUT: TECH1; stand; tech4; mod; sampstat; !user-specified output options;
```

Multiple Group Invariance (MGI1 in Table 2) With No Invariance Constraints

```
TITLE: Syntax for Model MGI1 (see Table 2; no invariance constraints)
DATA: FILE IS USCSET2YRGRPN=30444NV=52.DAT; !user-specified data file;
VARIABLE: NAMES ARE q1-q38 q40 year enroll rank per395
   flrn fenth forgn fgrp find fbrd fexam fasgn fwork;
   USEVARIABLES q1-q29 q30 q31 q32-q35; ! variables actually used in the analysis
  grouping is year (1=g1 2=g2); ! a dichotomous grouping variable;
ANALYSIS:
    ROTATION=GEOMIN(OBLIQUE, .5);
  MODEL:  f1-f9 BY q1-q35 (*1); ! '*1' means EFA;
    [f1-f9@0]; ! latent FMs fixed at 0 in G1;
    [q1-q35]; ! item intercepts free in G1 (default);
    q1-q35; ! item UNQ free in G1 (default);
  MODEL g2: f1-f9 BY q1-q35 (*1); !FLs free in G2;
    [q1-q35]; ! item intercepts free in G2 (default);
    q1-q35; ! item UNQ free in G2 (default);
  OUTPUT: TECH1; stand; tech4; mod; sampstat; !user-specified output options;
```

Multiple Group Invariance (MGI13 in Table 2) With Complete Invariance of All Parameters

Note that the "!" symbol is a comment so that text following !s (in italics) is ignored.

```
TITLE: Syntax for Model MGI13 (see Table 2; complete invariance)
DATA: FILE IS USCSET2YRGRPN=30444NV=52.DAT; !user-specified data file;
VARIABLE: NAMES ARE q1-q38 q40 year enroll rank per395
   flrn fenth forgn fgrp find fbrd fexam fasgn fwork;
   USEVARIABLES q1-q29 q30 q31 q32-q35; ! variables actually used in the analysis
  grouping is year (1=g1 2=g2); ! a dichotomous grouping variable;
ANALYSIS:
    ROTATION=GEOMIN(OBLIQUE, .5);
  MODEL: f1-f9 BY q1-q35 (*1); ! '*1' means f1 to f9 are ESEM factors;
    [Q1-Q35]; !intercepts free in G1 (default)
      !FACTOR VAR/COVAR (numbers in parentheses are user-defined parameter designations
          ! used to define invariance constraints in the multiple groups;
  f1 with f2 (1);f1 with f3 (2);f1 with f4 (3);f1 with f5 (4);
  f1 with f6 (5);f1 with f7 (6);f1 with f8 (7);f1 with f9 (8);
  f2 with f3 (9);f2 with f4 (10);f2 with f5 (11);
  f2 with f6 (12);f2 with f7 (13);f2 with f8 (14);f2 with f9 (15);
```

```
  f3 with f4 (16);f3 with f5 (17);
  f3 with f6 (18);f3 with f7 (19);f3 with f8 (20);f3 with f9 (21);
  f4 with f5 (22);f4 with f6 (23);f4 with f7 (24);f4 with f8 (25);f4 with f9 (26);
  f5 with f6 (27);f5 with f7 (28);f5 with f8 (29);f5 with f9 (30);
  f6 with f7 (31);f6 with f8 (32);f6 with f9 (33);
  f7 with f8 (34);f7 with f9 (35);
  f8 with f9 (36);
  F1-F9@1; ! FVARS fixed a 1 in G1 (defualt)
  !ITEM UNIQUENESSES (numbers in parentheses are user-defined parameter designations
        ! used to define invariance constraints in the multiple groups;
  q1(41); q2(42); q3(43); q4(44); q5(45);
  q6(46); q7(47);q8(48); q9(49); q10(50);
  q11(51); q12(52); q13(53); q14(54); q15(55);
  q16(56); q17(57);q18(58); q19(59); q20(60);
  q21(61); q22(62); q23(63); q24(64); q25(65);
  q26(66); q27(67);q28(68); q29(69); q30(70);
  q31(71); q32(72); q33(73); q34(74); q35(75);
  ! LATENT FMEANS (AL)
  [f1-f9@0]; ! LATENT FMEANS fixed at zero
      ! with item intercepts free unless testing FMean Invar
MODEL g2:
  !FACTOR LOADINGS (LAMBDA)
      !f1-f9 BY q1-q35 (*1); !By commenting out this line the model
        ! defaults to FL INV in G2; FL INV is all or none
  !ITEM INTERCEPTS (NU)
      ![Q1-Q35]; !By commenting out this line the model
        ! defaults to Intercepts invariant in G2;
      !FACTOR VAR/COVAR (numbers in parentheses are user-defined parameter designations
          ! used to define invariance constraints in the multiple groups;
    f1 with f2 (1);f1 with f3 (2);f1 with f4 (3);f1 with f5 (4);
    f1 with f6 (5);f1 with f7 (6);f1 with f8 (7);f1 with f9 (8);
    f2 with f3 (9);f2 with f4 (10);f2 with f5 (11);
    f2 with f6 (12);f2 with f7 (13);f2 with f8 (14);f2 with f9 (15);
    f3 with f4 (16);f3 with f5 (17);
    f3 with f6 (18);f3 with f7 (19);f3 with f8 (20);f3 with f9 (21);
    f4 with f5 (22);f4 with f6 (23);f4 with f7 (24);f4 with f8 (25);f4 with f9 (26);
    f5 with f6 (27);f5 with f7 (28);f5 with f8 (29);f5 with f9 (30);
    f6 with f7 (31);f6 with f8 (32);f6 with f9 (33);
    f7 with f8 (34);f7 with f9 (35);
    f8 with f9 (36);
    F1-F9@1; !FVARS fixed a 1 in G2 (default) but must specify when FCOV Invariant
  !ITEM UNIQUENESSES (numbers in parentheses are user-defined parameter designations
          ! used to define invariance constraints in the multiple groups;
    q1(41); q2(42); q3(43); q4(44); q5(45);
    q6(46); q7(47);q8(48); q9(49); q10(50);
    q11(51); q12(52); q13(53); q14(54); q15(55);
    q16(56); q17(57);q18(58); q19(59); q20(60);
    q21(61); q22(62); q23(63); q24(64); q25(65);
    q26(66); q27(67);q28(68); q29(69); q30(70);
    q31(71); q32(72); q33(73); q34(74); q35(75);
    ! LATENT FMEANS
    [f1-f9@0];
     ! LATENT FMEANS fixed at zero with inter invar set free unless testing FMean Invar
  OUTPUT: TECH1; stand; tech4; mod; sampstat; !user-specified output options;
```

MIMIC Latent Growth Model (MIMIC-GRW1A in Table 2) Linear, Quad, and Cubic Growth Components

```
TITLE: Syntax for Latent Growth Model MIMIC-GRW1A (see Table 2)
DATA: FILE IS uscset2YRgrpN=30444NV=55.dat; !user-specified data file;
VARIABLE: NAMES ARE q1-q38 q40 year enroll rank per395
   flrn fenth forgn fgrp find fbrd fexam fasgn fwork YRLin YRQud YRCub;
   USEVARIABLES q1-q29 q30 q31 q32-q35 YRLin YRQud YRCub;
       ! variables actually used in the analysis
  ANALYSIS:
     ROTATION=GEOMIN(OBLIQUE, .5);
  MODEL: f1-f9 BY q1-q35 (*1); ! '*1' means f1 to f9 are ESEM factors;
  f1-f9 on YRLin YRQud YRCub; ! polynomial growth components regressed on 9 ESEM factors;
  OUTPUT: TECH1; stand; tech4; tech5; mod; sampstat;
```

MIMIC Model of Background (Linear and Quadratic) Effects (MIMIC-BCK2A in Table 2)

```
TITLE: Syntax for MIMIC Model for Background Effects MIMIC-BCK2A (see Table 2)
DATA: FILE IS uscset2YRgrpN=30444NV=61.dat; !user-specified data file;
VARIABLE: NAMES ARE q1-q38 q40 grp2 enroll rank per395
     flrn fenth forgn fgrp find fbrd fexam fasgn fwork
     zq38 zq36 zenr renrsq rq38sq rq36sq;
     USEVARIABLES Q1-Q35 zq36 zenr zq38 renrsq rq38sq rq36sq;
       ! variables actually used in the analysis
ANALYSIS:
    ROTATION=GEOMIN(OBLIQUE, .5);
MODEL: f1-f9 BY q1-q35 (*1); ! '*1' means f1 to f9 are ESEM factors;
    f1-f9 on zq36-rq36sq;
  ! Linear and quad effects of three background variables regressed on 9 ESEM factors;
OUTPUT: TECH1; stdyx; mod; sampstat;
```

Higher Order Factor Model (HO1A in Table 2)

```
TITLE: Syntax for Higher-order Factor Model (HO1A in Table 2)
DATA: FILE IS USCSET2YRGRPN=30444NV=52.DAT; !user-specified data file;
VARIABLE: NAMES ARE q1-q38 q40 year enroll rank per395
    flrn fenth forgn fgrp find fbrd fexam fasgn fwork;
  USEVARIABLES q1-q29 q30 q31 q32-q35; ! variables actually used in the analysis;
ANALYSIS:
  ROTATION=GEOMIN(OBLIQUE, .5);
MODEL: f1-f9 BY q1-q29 q32-q35 (*1);
    ! 9 ESEM factors are defined without the overall rating items Q31 and Q32;
  f10 by q30 q31; !a CFA factor is defined by the two overall rating items Q31 and Q32;
  f1-f9 on f10; ! global overall rating factor regressed on 9 ESEM factors
OUTPUT: TECH1; stand; tech4; mod; sampstat;
```

For more details on M*plus* syntax, see the M*plus* Users Manual (Muthén & Muthén, 2008). Also see http://www.statmodel.com/ugexcerpts.shtml for downloadable copies of the Users Manual, the M*plus* 5.1 Language Addendum, and the M*plus* 5.1 Examples Addendum (which give annotated examples and the modeling steps for other applications).