Running Head: Shape and Level

Disentangling Shape from Levels Effects in Person-Centered Analyses: An Illustration Based University Teacher Multidimensional Profiles of Effectiveness.

Alexandre J.S. Morin
University of Western Sydney (Australia)
Herbert W. Marsh
University of Western Sydney (Australia), University of Oxford (UK), & King Saud University (Saudi Arabia)

This is a prepublication version of a manuscript to be published by *Structural Equation Modeling*:
Morin, A.J.S., & Marsh, H.W. (2015). Disentangling Shape from Levels Effects in Person-Centred Analyses: An Illustration Based University Teacher Multidimensional Profiles of Effectiveness. *Structural Equation Modeling*, *22* (1), 39-59.
Accepted on 6 August 2013.

Corresponding author:
Alexandre J.S. Morin
Centre for Positive Psychology and Education
University of Western Sydney
Locked Bag 1797, Penrith, NSW 2751, Australia
E-mail: A.Morin@uws.edu.au

Abstract

This study compares alternative ways of disentangling the effects of *levels* (the tendency for a person to be high, medium or low across all factors in the profile) and *shape* (the tendency for a person to have a distinct pattern of factors on which they are high, medium or low) in profile analyses. This issue is particularly relevant to performance appraisals where it is often useful to identify specific strengths and weaknesses over and above a person global performance, but also to person-centered analyses more generally where the observation of qualitative (*shape*) differences between profiles is often used as justification for the added-value of profiles. Substantively, this study illustrates these issues in the identification of profiles of teachers based on multidimensional students' ratings of their effectiveness, using an archival data set of 31,951 class-average ratings based on the Students' Evaluations of Educational Quality (SEEQ) instrument collected over a 13-year period. The results show the superiority of a factor mixture operationalization of teaching effectiveness in which a global effectiveness factor was used to control for unnecessary *level* effects in the profiles.

Key words: Latent Profile Analyses, Factor Mixture Analyses, Shape, Level, Performance Evaluation, Students' Evaluations, University Teaching, Teacher profiles.

Complex substantive issues often require sophisticated methodologies, and methodological insights may also emerge from attempts to answer complex substantive problems—this is the essence of methodological-substantive synergies (Marsh & Hau, 2007). Methodological-substantive synergies are joint ventures in which new methodological developments are applied to, or emerge from, substantively important issues. Methodologically, this study contrasts alternative ways of disentangling the effects of *levels* (the tendency for a given person to be high, medium or low across all factors in the profile) and *shape* (the tendency for a given person to have a distinct pattern of factors on which they are high, medium or low) in order to maximize the meaningfulness and practical utility of profiles. For instance, this issue is central to the study of profiles of competencies in the context of performance appraisals conducted for developmental purposes as it allows for an identification of the specific strengths and weaknesses over and above the global performance of a person across indicators. Overall, this issue has broad relevance for the person-centered investigation of profiles based on multidimensional constructs characterized by a combination of global and domain-specific components, such as self-concept (e.g. Marsh, 2007a), commitment (Morin, Morizot, Boudrias, & Madore, 2011), etc. Substantively, this study illustrates these issues in the identification of profiles of teachers based on multidimensional students' ratings of their effectiveness.

Methodological Issues: Disentangling Shape and Level Effects in Mixture Models.

A Person-Centered Perspective on Performance Evaluation

Person-centered approaches (Bergman, Magnusson, & El-Khouri, 2003; Bergman & Trost, 2006) date back to Allport (1937), Myers and McCaulley (1985), and even to Greek philosophers, who first observed that humans need to classify objects to better make sense of their surroundings. Taxonomies, or typologies, are classification systems designed to help categorize objects/individuals more accurately into qualitatively and quantitatively distinct subgroups or profiles (Bailey, 1994; Bergman et al., 2003). Unfortunately, these approaches have yet to be systematically incorporated into research on performance evaluation. Indeed, although the practice of performance evaluation often explicitly aims to identify distinct profiles of individuals based on multidimensional ratings of competencies (e.g., Hobson & Gibson, 1983; Swank, Taylor, Brady, & Freiberg, 1989), the research supporting this field of practice is traditionally anchored in variable-centered analyses (e g. regression, factor analysis) (see, e.g., Fletcher, 2001; Latham & Mann, 2006). The results from variable-centered studies represent a synthesis (or averaged estimate) of the relations observed in every individual from the sample under study, without systematically considering that these relations may differ across subgroups of participants. Variable-centered methods end up summarizing data by average levels and variability in different dimensions of competencies, across observed subgroups or measurement points. Conversely, person-centered analyses generate a typology in which participants are classified into qualitatively and quantitatively distinct profiles based on their specific combinations of strengths and weaknesses on the same array of competencies. A teaching effectiveness typology would thus classify teachers into groups so that those within a group have a similar configuration of skills (e.g., strong on organization and evaluation, but weaker in managing group interaction), while displaying a profile that is qualitatively and quantitatively distinct from other groups' profiles.

Recent technological developments (e.g., Muthén & Muthén, 1998-2010; Vermunt, & Magidson, 2000), and user-friendly introductions (e.g., Muthén, 2002; Vermunt & Magidson, 2002) have brought mixture modeling methods (McLachlan, & Peel, 2000; Muthén & Shedden, 1999) into mainstream psychological and educational research where they have superseded – perhaps due to their greater flexibility – cluster analytic (e.g., Magidson, & Vermunt, 2002; Vermunt, 2011; but also see Steinley & Brusco, 2011) and taxometric methods (e.g., Lubke & Tueller, 2010; Waller & Meehl, 1998). The key difference between person-centered mixture models (e.g., latent profile analysis– LPA) and variable-centered factor analyses is the nature of the estimated latent variable: categorical in the first case and continuous in the second case. Thus, "the common factor model decomposes the covariances

to highlight relationships among the variables, whereas the latent profile model decomposes the covariances to highlight relationships among individuals" (Bauer & Curran, 2004, p. 6). Thus, factor models regroup variables, whereas LPAs regroup persons (Cattel, 1952; Lubke & Muthén, 2005).

Choosing Between Person-Centered and Variable-Centered Representations

Choosing between these alternative representations is not easy since a *k*-class LPA model has identical covariance implications than a *k-1* common factor model and thus represents an equivalent model (Bauer & Curran, 2004; Steinley & McDonald, 2007). Simulation studies also showed that spurious latent classes may emerge when none exist as a way to account for violations of the model distributional assumptions (e.g., Bauer, 2007). Multiple partial answers to this dilemma have been attempted (Lubke & Neale, 2006, 2008; Muthén, & Asparouhov, 2009; Steinley & McDonald, 2007). However, the existence of statistical models presenting equivalent approximation of the data but providing radically different explanations of the reality is almost universal in the social sciences (Cudeck & Henly, 2003; Hershberger, 2006; Muthén, 2003). In the end, the best way to support a substantive interpretation of the profiles as reflecting significant subgroups is to embark on a process of construct validation, including an assessment of the heuristic value and theoretical conformity of the profiles, as well as tests of their generalizabiltiy to new samples (Cudeck & Henly, 2003; Marsh, Lüdtke, Trautwein, & Morin, 2009; Morin, Morizot et al., 2011; Muthén, 2003).

In the midst of this debate, one criterion that has [mostly] always been implicit in the person-centered literature – and was even explicitly mentioned by some (e.g., Bauer, 2007; De Boek, Wilson, & Acton, 2005) – is the need to observe qualitative (*shape*) differences between the extracted profiles in order to support their meaningfulness. The main argument supporting this assumption is that *ordered* profiles, showing only quantitative *level* differences (i.e., with one profile simply presenting a higher level than the other on the variables considered), would be better represented by variable-centered methodologies, and would thus have no heuristic value. Although Muthén (2001, p. 8) argued that "with ordered classes, one may ask what advantage LCA [latent class analysis] has versus doing regular factor analysis […]. The answer is that LCA helps find cluster of individuals who are similar, whereas this is difficult in factor analysis". This argument is similar to Nagin (2010, p.61) affirmation that profiles differing only quantitatively (i.e., showing only *level* differences) may serve to represent a nonlinear distribution by a finite number of "point of support". However, arguing that extracted latent profiles simply serve to better represent complex non-linear relationships is only an alternative way of saying that non-linear variable-centered analyses would offer a better representation of the variables (also see Bauer & Shanahan, 2007). Indeed, Muthén (2001, 2006; Muthén & Asparouhov, 2006), and others (Krueger, Markon, Patrick, & Iacono, 2005; Kuo, Aggen, Prescott, Kendler, & Neale, 2008), previously used the observation of ordered profiles as an argument to change the specification of the model by including a latent continuous factor in conjunction with a latent categorical variable (i.e., a factor mixture model, e.g., Lubke & Muthén, 2005), to obtain cleaner *shape* differences. However, we reinforce that the strongest test of the meaningfulness of extracted profiles has to do with their correspondence to theoretical expectations. For this reason, we caution readers against the use of suboptimal two-step approaches in which latent profiles are first extracted based on theoretical expectations and, upon observing that they show only level differences, new models designed to disentangle *shape* from *level* differences are implemented as a way to salvage an otherwise meaningless solution. In such cases, parsimony would rather dictate that a variable-centred common factor model be pursued as the best representation of pure *level* effects. Although we propose different approaches in order to disentangle *shape* and *level* differences in latent profile models, we argue that these models should be anchored into a clear theoretical rationale showing that both *shape* and *level* effects can be expected to be substantively meaningful.

Shape Differences as a Prerequisite to Person-Centered Analyses

Thus, although the question as to whether extracted latent profiles of participants really do

reflect meaningful subgroups of individuals is a complex issue, the need to observe qualitative, *shape* differences between the extracted profiles does seem to reflect an important prerequisite. Counter-examples to this implicit rule are indeed very hard to locate in the published literature, and generally still show shape-related differences on at least some of the parameters freely estimated across profiles (e.g., Morin, Rodriguez, Fallu, Maïano, & Janosz, 2012). When using effectiveness evaluations for developmental purposes, it is particularly important to distinguish between the *level* of a profile (for example, whether the ratings across all the effectiveness scales are consistently high or low for a particular person) and the *shape* of a profile (for example, whether each person has a distinguishable profile of scores characterized by specific areas of strengths and weaknesses). Even before current technological developments, this distinction between *level* and *shape* was considered as a main objective of repeated measures multivariate analyses of variance – MANOVAs. For instance, in an important precursor of the present study, Marsh and Bailey (1993) proposed that repeated measures MANOVAS could be used to investigate whether individuals differed from one another simply on their overall level of effectiveness (i.e., *level* effects) or whether profiles of individuals presenting different patterns of competencies (i.e., *shape* effects) were also present.

More precisely, this method implies that repeated multidimensional ratings of competencies (or any multidimensional construct) be available for a sample of individuals. Then, the multiple dimensions are treated as repeated measures with the individuals as the grouping variable and the time-specific assessments (grouped within persons) as the basic unit of analysis. From this analysis, any main effect of individuals on the ratings is indicative of *level* effects, whereas any interaction effect between individuals and the dimensions is indicative of *shape* effects. Thus, when a sufficient proportion of variability can be attributed to *shape* effects, this suggests that person-centered analyses may be appropriate to pursue. Otherwise, variable-centered analyses are likely sufficient and person-centered analyses might actually prove suboptimal. However, when the results show the presence of strong *level* and *shape* effects, then models allowing for the partialling out of both facets might be a worthy alternative to consider. We leave open the question as to what represent a sufficiently large proportion of the variability to justify the consideration of models allowing for the analysis of *level*, *shape*, or dual *shape* and *level* effects in the data. However, results from this MANOVA-based approach are routinely accompanied by various multivariate effect sizes indicators (e.g., eta^2) that can be interpreted in line with Cohen (1988) guidelines as to what represents a small, moderate, or large effect size. Unless substantive domain-specific guidelines suggest otherwise, we suggest that the effect size associated with either *level* and/or *shape* effects should be at least moderate in magnitude to justify their consideration in the analyses. Obviously, we do not propose this MANOVA method as a necessary first step to the conduct of person-centered analyses, as this would be an unrealistic expectation for most research where repeated measures are not available. However, when possible, this preliminary test provides a strong test of whether shape effects are present in the multidimensional ratings – justifying the use of person-centered analyses –, and whether *level* effects are also strong enough to justify considering methods allowing for the separation of *shape* and *level* effects in the analyses. In other situations, researchers will need to base these decisions on previous research results, theoretical frameworks, and substantive a priori expectations.

Partialling out Level Effects for Clearer Shape Differences: Four Alternative Models

In cases where both *level* and *shape* effects are expected to be strong, the identification of qualitatively distinct profiles becomes harder since strong *level* effects tend to create equally strong quantitative differences. For instance, some profiles may include generally strong teachers when compared to other teachers. However these profiles may still present relative areas of strength and weaknesses worthy of consideration in relation to specific dimensions, although these may be harder to identify given that even these weaknesses may be relatively strong compared to the levels observed in weaker profiles. Alternative specifications of mixture models may be used to tackle this issue.

A classical LPA model is presented in the Figure 1-Model 1. LPA postulates that the correlations between the dimensions, or profile indicators, may be explained by the presence of a categorical latent variable representing distinct profiles of individuals. The use of such a model should be based on the expectation that *shape* and *level* differences do not need to be disentangled from one another, or that there is no reason to expect *level* differences in the extraction of the profiles. An alternative LPA model is present in Figure 1-Model 2 and specifically includes a higher-order dimension (estimated from the covariance among the first-order dimensions) designed to explicitly reflect *level* effects in the extracted latent profiles (for a related discussion, see Marsh, Lüdtke et al., 2009). In contrast with the previous one, this model assumes that *level* differences need to be taken into account in the interpretation of the profiles, but are unlikely to hide meaningful *shape* differences between the profiles. However, this model is unlikely to provide a solution to the *shape-level* dilemma when strong effects of *level* in the definition of the profile are expected as it directly allows *level* effects, as represented by the higher order dimension, to influence the classification. Thus, this model is likely to result in even stronger *level* differences between the profiles. By default, LPA assumes conditional independence: conditional on class membership, the residual correlations between the observed variables should be zero (e.g., Vermunt & Magidson, 2002). In other words, the latent profiles are assumed to explain all of the correlations between dimensions. However, this assumption is often too stringent with real-life data, especially when the research question does not necessarily assume conditional independence, such as when strong level effects are known to be present (Vermunt & Magidson, 2002; Uebersax, 1999). In such cases, spurious latent classes may even emerge as a way to reconcile the data with these unrealistic assumptions (Bauer, 2007).

Factor mixture analyses (FMA) were proposed as a way to solve this issue and to extract, by way of a continuous latent factor, the *level* variance that is shared by the dimensions (Lubke & Muthén, 2005; also see Masyn, Henderson, & Greenbaum, 2010). FMAs thus represent an efficient way of including correlations between the indicators by allowing them to simultaneously relate to a categorical latent variable (the LPA model) and to a continuous latent variable (the common factor model). This method allows for conditional dependence among the indicators in a more parsimonious way (i.e., with fewer parameters) than the alternative of directly specifying correlations among the indicators' residuals (e.g., Uebersax, 1999). Figure 1-Model 3 presents a model previously described by Morin, Morizot et al. (2011) as a way to represent higher-order *level* effects through the inclusion of a class-invariant continuous latent factor. Thus, in this model, the covariance between the full set of effectiveness dimensions is used to define a higher-order continuous latent factor designed to explicitly reflect *level* effects (i.e., overall level of effectiveness) in the extracted latent profiles while the covariance left unexplained by this common factor is used to estimate the latent categorical variable representing the profiles. This model is thus similar to a bifactor model where the specific "factors" would in fact be categorical and reflect the profiles. More precisely, a bifactor model (Holzinger & Swineford, 1937; Morin, Tran, & Caci, 2013; Reise, 2012) analyses the total covariance among the indicators to extract a global G factor underlying all indicators, and models the residual covariance not explained by the G factor through the specific S factors. According to Chen, West and Sousa (2006, p.190): "Bifactor models are potentially applicable when (a) there is a general factor that is hypothesized to account for the commonality of the items; (b) there are multiple domain specific factors, each of which is hypothesized to account for the unique influence of the specific domain over and above the general factor; and (c) researchers may be interested in the domain specific factors as well as the common factor that is of focal interest."

Model 3 is perfectly in line with this operationalization and relies on the assumption that strong *level* effects would be present in the data due to the presence of a substantively meaningful global continuous construct underlying all of the dimensions considered and that this continuous latent construct has a meaning in and of itself. This model further assumes that meaningful specific *shape*-

based profiles would emerge over-and-above this continuous latent factor and are themselves deserving of being taken into account. In the current application, we argue that teachers differ from one another on the basis of some global competency indicator that needs to be considered in and of itself in assessing their teaching effectiveness. We also argue that over and above this overall level of effectiveness they also present specific profiles of strengths and weaknesses that also need to be taken into account, making this model the most theoretically suitable to the current substantive application.

Although others have proposed similar models as way to obtain sharper qualitative difference between profiles (e.g., Kuo et al., 2008; Muthén & Asparouhov, 2006), they relied on continuous factors specified as totally or partially non-invariant across the latent classes, inducing confusion in the results. Indeed, in mixture models, the latent categorical variable depicting the profiles is always estimated on the basis of all parameters left non-invariant across classes. In other words, a FMA with a class-varying continuous factor becomes a way of probing for the non-invariance of the common factor model across the latent profiles (e.g., Tay, Newman, & Vermunt, 2011). Such a model could thus result in profiles including participants with the same shape and levels on the various dimensions being assessed, but differing in the way the common factor underlying these dimensions is specified (i.e., different loadings, uniquenesses, etc.). To make matters worse, when the profiles start to differ on the nature of the common factor, they also cease to be directly comparable on the main constructs of interest (i.e., dimensions of effectiveness). Indeed, the profiles themselves are defined from the part of these constructs that is left unexplained by the common factor. Thus, if the measurement model underlying this common factor changes from one profile to the other, then the part of the effectiveness dimension that is not explained by this common factor also cease to be comparable across profiles. In other words, in a model in which teachers' global effectiveness is extracted from ratings of specific competency indicators in order to estimate clearer profiles, then it is important that the way "global effectiveness" is defined be the same for all teachers. This does not mean that FMA with a class-varying factor structure are not useful. Indeed these models are likely the most appropriate way to ensure that a measurement model is fully invariant across all possible subpopulations forming a sample and to the most stringent test of measurement bias that can be conducted psychometrically. However, these models are not appropriate when the objective is to simply partial out level effects in order to estimate clearer shape-differences in profiles of participants. In this case, we argue that the common factor model that is part of the FMA should be specified as invariant across profiles.

A possible limitation of Model 3 is that all parts of the model are simultaneously estimated. Thus, whereas the profiles are estimated from the part of the dimensions that remains unexplained by the continuous common factor, so is the continuous common factor estimated from the part of the dimensions that remain unexplained by the profiles. Thus, the continuous common factor is estimated so as to reflect the global *level* of effectiveness, but only from the part of this overall *level* that is not better explained by the categorical latent variable representing the profiles. This means that some part of this global *level* of effectiveness may remain a part of the profiles to create quantitative differences between them. To address this issue, Figure 1-Model 4 proposes to include the higher-order dimension (estimated from the covariance among the first-order dimensions) as a controlled variable in the model so as to estimate profiles based on purer *shape*-related qualitative differences. This higher order dimension is related to the first-order dimensions through regressions, rather than factor loadings. Thus, the resulting profiles are estimated from the residuals of these predictions (when predictions are fixed as invariant between classes). This model thus allow for the estimation of profiles based on effectiveness dimensions that are centered at the mean of the global effectiveness factor estimated in each of the extracted profiles (i.e., group-mean centered), excluding any form of level-differences. Although it appears similar to the previous model, this model relies on highly different substantive assumptions. Indeed, substantively, this model addresses the situation where global *level* effects are seen as some form of biasing influence (i.e. halo effect, social desirability, shared method variance)

that have no substantive meaning in and of themselves, and that need to be controlled for before the extraction of the latent profiles. However, this forced extraction of all *level* effects from the estimation process may, when there is reason to expect that the global *level* effects are meaningful, result into less "natural" profiles that are harder to connect to the real-life reality of the individuals under evaluation.

It should be noted that all of these models allow for the possibility to control additional variables known to result in level effects as direct predictors of the dimensions in the same manner as the higher-order dimension from Model 4. For instance, we know that teacher effectiveness based in students' ratings tend to be substantially higher in graduate classes than in undergraduate classes (see Marsh, 2007b; Marsh & Bailey, 1993), suggesting that this variable should be controlled in analyses aiming to partial out level effects from the estimation of teachers' profiles of effectiveness.

Finally, we reinforce that we do not propose these models a component of routine applications of person-centered analyses. We rather propose them as alternatives allowing for the separation of *shape* from *level* effects in the estimations of latent profiles when there are strong theoretical or empirical reasons to expect this to be necessary. In fact, we further argue that an important pre-requisite to the use of these models should be the presence of clear empirical and/or theoretical a priori favouring one model above the others. Although this study is mainly methodological, we illustrate these issues based on a real dataset including university students' evaluations of teaching effectiveness (SETs). This substantive area was selected as particularly well-suited to the issues considered in this paper. However, as noted above, one of the main criteria against which to evaluate the meaningfulness of extracted profiles is their conformity with theoretical expectations. Similarly, a main difference between the four proposed models has to do with their substantive implications. For these reasons, we now move to a short substantive introduction to the SETs literature that is most relevant to this application. Methodologically-oriented readers may feel free to skip the next sections.

Substantive Application: Students' evaluations of teaching effectiveness (SETs).

Multiple Dimensions of Teaching Effectiveness

As an inherently complex activity teaching comprises multiple interrelated components (e.g., clarity, organization, enthusiasm) that should to be simultaneously considered when evaluating teaching quality (Feldman, 1997; Marsh, 2007b; Marsh & Roche, 1993; Renaud & Murray, 2005). Since SETs are generally specifically designed as formative feedback tools intended to contribute to the improvement of teaching, their multidimensionality is especially important in order to target specific areas of improvement. Strong support for the multidimensionality of SETs comes from the Students' Evaluations of Educational Quality (SEEQ) instrument (Marsh, 1982; 1987; 2007b; Marsh & Hocevar, 1991; Richardson, 2005). The SEEQ assesses nine factors are assessed on a five point answer scale referring to the teacher and ranging from very poor to very good: 1- Learning/Value (i.e., the course was valuable learning experience, was intellectually stimulating/challenging); 2 - Enthusiasm (i.e., the instructor displayed enthusiasm, energy, and ability to hold interest); 3- Organization/Clarity (i.e., the quality of organization and clarity of the explanations, materials, objectives, and lectures); 4- Group Interaction (i.e., students were encouraged to participate, share ideas and ask questions); 5- Individual Rapport (i.e., the instructor is accessible, and interested in students); 6- Breadth of Coverage (i.e., the courses includes the presentation of background, concepts and alternative approaches or theories); 7- Exam/Grading (i.e., perceived value and fairness of the exams and grading); 8- Readings/Assignments (i.e., perceived value of assignments in adding appreciation and understanding); 9- Workload/Difficulty (i.e., perceived difficulty, workload, pace, and hours outside of class). The factor structure of SEEQ has been replicated in many published studies, but the most compelling support is provided by Marsh and Hocevar (1991) who replicated this structure in 21 different groups of classes differing in terms of course level, instructor rank, and academic discipline on an archive of more than 40,000 sets of class-average ratings.

Potential Profiles of Teacher Effectiveness

Among the few previous cluster analytic studies which attempted to describe teacher's profiles on diverse characteristics, apparently none focused on University teachers and most focused on characteristics not directly relevant to the effectiveness of their teaching, such as professional identity, motivation, personal learning or change adoption practices (e.g., Canrinus, Helms-Lorenz, Beijaard, Buitink, & Hoffman, 2011; Oscarson & Finch, 1979; Pedder, 2007; Wang & Liu, 2008). Fortunately, a least some studies sought to identify teacher's profiles on based on characteristics more directly relevant to teaching effectiveness at the primary of secondary school level. However, most of these studies focused on a limited set of two or three variables, neglecting to take into account the full multidimensionality of teaching effectiveness (e.g. Brekelmans, Levy, & Rodriguez, 1993; Brok, Fisher, Brekelmans, Wubbels, & Rickards, 2006; Rickards, Brok, & Fisher, 2005) and yielded results that could be interpreted as showing mostly *level* differences. Similarly, contrasting teachers traditional (teacher-centered) versus constructivist (student-centered) beliefs in a sample of Chinese primary school teachers, Sang, Valcke, Braak, and Tondeur (2009) identified four distinct profiles reflecting different combinations of high or low *levels* of both beliefs. Prawat (1985) reached similar conclusion by contrasting American elementary school teachers' beliefs regarding the relative priority they attribute to the cognitive versus affective development of children. James and Pedder (2006) studied 558 primary and secondary school UK teachers regarding their beliefs and practices regarding classroom assessment practices. Their results revealed five different clusters showing *shape* and *level* differences. Although these results looked promising in regards to the identification of *shape* effects, an attempt to replicate them by Winterbottom, Taber, Brindley, Fisher, Finney, and Riga (2008) failed to do so and only identified 4 clusters differing mainly on *level*. Finally, in an extensive observational study using 30 indices of middle school teachers effectiveness, Swank et al. (1989) converged on a three cluster solution that they explicitly interpreted as showing only quantitative differences regarding teachers' overall *levels* of effectiveness. Clearly, these studies attest to the presence of strong *level* effects in multidimensional assessments of teaching effectiveness and leave open the question of whether meaningful *shape*-based differences could be identified when the assessment is based on reliable multidimensional instruments such as the SEEQ.

Given the lack of research on profiles of teachers' multidimensional SETs, it is hard to propose clear hypotheses regarding the nature of the expected shape-based profiles. However, some general theoretical models of human identity and teaching, and some implicit assumptions used in previous studies of teachers' profiles allow us to formulate some expectations. Indeed, multiple theories of human identity emphasize the presence of two opposite tendencies that may likely impact teaching style, one centered on interpersonal relations, and one centered on autonomy and achievement (e.g., Brewer, 1991; Cross, & Madson, 1997; Helgeson, 1994; McClelland, 1987). Parallel distinction have been made regarding teachers values and practices based on the importance attributed to contributing to the affective development of the students in addition to their cognitive development (Prawat, 1985; Shavelson & Stern, 1981; Wooley, Benjamin, & Wooley, 2004), and on the level of control left to students in the classroom (Brekelmans et al., 1993; Brok et al., 2006). Consistent with a growing body of evidence showing that teachers beliefs do indeed affect teaching practices and students outcomes (Meece, Anderman, & Anderman, 2005; Roseth, Johnson, & Johnson, 2008; Shavelson & Stern, 1981; Wooley et al., 2004), we expect that some profiles will differ according to the importance attributed to *affective* relations with students (Enthusiasm, Group Interaction, Individual Rapport), versus simply ensuring that their learning experiences are *cognitively* complete (Exam/Grading, Learning/Value, Organization/Clarity, Readings/Assignments, Breadth of Coverage, Workload/Difficulty). It is interesting to note that this distinction parallels an illustration of possible profiles provided by Marsh & Bailey (1993). Additional distinctions have been proposed between formative and summative assessments (Black, McCormick, James, & Pedder, 2006; James & Pedder, 2006) and between performance-ability (where students are asked to demonstrate their relative abilities and to "be the

best) versus mastery (where students are encouraged to master course content and develop competencies) goal practices (Ames, 1992; Meece et al., 2005; Midgley, 2002; Roseth et al., 2008). We can thus also expect to observe profile differences with higher levels of Learning/Value, Organization/Clarity, Readings/Assignments, and Breadth of Coverage for mastery-oriented teachers and higher levels on Exam/Grading and Workload/Difficulty for performance-oriented teachers.

Temporal Stability of Teachers Profiles as an Indicator of their Construct-Validity

In a classic study particularly relevant to the present investigation, Marsh and Bailey (1993, also see Hativa, 1996) used repeated measures (M)ANOVAs on the 9 SEEQ scores (treated as the repeated measure) on a sample of 123 teachers (treated as the grouping variable) who had been evaluated repeatedly over a 13 years period by a total of 3079 classes. Their results showed strong (e.g. Cohen, 1988) *level* effects, explaining 37% of the variability (according to the eta^2 indicator) in SEEQ ratings, consistent with longitudinal stability of overall teaching effectiveness and with the previous results. However, they also found an even stronger interaction effect between teachers and factors showing that 47% of the SEEQ ratings are due to the presence of stable *shape*, or profiles, effects. Interestingly, these profiles generalized across subject and course level. Clearly, the existence of stable profiles has important implications for feedback interventions and most importantly for the understanding of teaching effectiveness. Marsh and Bailey (1993) results strongly suggest the presence of substantively meaningful shape and level effects in SETs (corresponding to Model 3). This is also consistent with the practice of performance evaluation and feedback where one is usually interested in both the global effectiveness as well as the more specific profile of strength and weaknesses of a person. In other words, we would expect profiles of teaching effectiveness to be best represented by Model 3, consistent with an interpretations of individual teachers' effectiveness based on both their global level of effectiveness and their more specific profiles of strength and weaknesses.

The issue of the stability of profiles is extremely important to person-centered analyses more generally and to performance appraisals more specifically, although very seldom investigated. Indeed, implicit to person-centered analyses and performance appraisals, is the assumption that the profiles are a function of the persons that are evaluated rather than, or in addition to, the situation (e.g., Fletcher, 2001; Reb & Greguras, 2010). The assumption that it is possible to identify stable profiles of individuals that can be used to guide selection, promotion, and other managerial decisions, but also to guide interventions allowing individuals to improve their profiles of competencies also lie at the core of performance appraisal practices (e.g., Reb & Greguras, 2010). We previously noted that the strongest support for an interpretation of profiles as reflecting significant subgroups is to embark on a process of construct validation (e.g., Marsh, Lüdtke et al., 2009; Muthén, 2003). Although longitudinal information is seldom available regarding the stability of profiles, we argue that demonstration of their stability likely represents one of the strongest tests of their construct validity. Interestingly, Marsh and Bailey (1993) results suggests that at least 47% of the variability in teachers multidimensional SETs can be expected to reflect stable shape-related profiles differences whereas 37% of the variability can be expected to reflect stable levels of effectiveness that generalize across dimensions.

The Present Study

The present study is methodological-substantive synergy in which we illustrate the use of LPA and FMA models for the identification of distinct profiles based on multidimensional ratings of effectiveness. More from precisely, we contrast the use of four alternative parameterizations of mixture models designed to partial out the effects of *levels* (the tendency for a given person to be stronger or weaker across all dimensions of effectiveness) profiles in order to obtain cleaner *shape*-based differences (illustrated by a distinct pattern of strengths and weaknesses). These issues of broad relevance to the field of performance evaluation and to person-centered research more generally are illustrated from the standpoint of teaching effectiveness research. Indeed, previous results attest to the presence of strong *level* effects in multidimensional students' ratings of teachers' effectiveness that

need to partialled out in order to obtain meaningful shape-related profiles.

Method

Sample and Procedure

Data come from an archive of SETs based on the SEEQ instrument (Marsh, 1982; Marsh & Bailey, 1993; Marsh, Muthén et al., 2009). 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 U.S. For purposes of the present investigation, 31,951 class-average sets of rating based on responses by at least 10 students, including all undergraduate and graduate level courses taught by regular faculty. This seminal archival data set includes ratings that have already been cleaned-up for multivariate outliers, inconsistent responding, and highly influential observations. In addition, being based on class average ratings, the likely impact of extreme individual cases was also much reduced to begin with. However, we note here the importance of conducting such preliminary verification, even using easy to use modern resources for applied research (see for instance Sterba & Peck, 2012) as extreme cases are likely to exert a substantial influence in mixture modeling contexts, even resulting in the potential extraction of small outlying classes. Within this larger data set, a total of 195 teachers were consistently evaluated over time, providing a total of 6025 class-average ratings (each having been rated by 16 to 61 different classes, with a mean of 30.9 classes). These teachers will be used to verify the stability of the profiles. Typically the SEEQ was distributed to students shortly before the end of academic terms, administered by a student or administrative staff according to standardized instructions, and taken to a central office where they were processed. Although participation was voluntary, the university required that all units collected some form of SETs and did not consider any personnel (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. Given the nested longitudinal subsample of teachers, this dataset was selected as particularly well-suited to the illustration of the methodological issues that are the object of this study. Indeed this subsample allowed us to investigate the longitudinal stability of the identified profiles, in addition to providing an initial test of the generalizability of the results over time. Further ensuring that the results could be expected to show some substantive generalizability, the extensive set of published results based on this university (e.g., Marsh, 1982; 1987; Marsh & Roche, 2000) are broadly consistent with findings from other SET research (for reviews, see Marsh, 1987; 2007b).

Analyses

All analyses are based on the nine standardized SEEQ factor scores (M = 0, SD = 1) obtained from the Exploratory Structural Equation Model (ESEM) recently reported by Marsh, Muthén et al. (2009; who also present extensive literature evidence supporting the decision to rely on an ESEM versus CFA structure for this instrument) and estimated on the same archival data set. The higher-order teacher effectiveness factor used in models 2 and 4 was also estimated starting from Marsh, Muthén et al. (2009) ESEM model converted to the ESEM-within-CFA framework in order to allow for the estimation of the higher-order factor (for a description of this method see Morin, Marsh, & Nagengast, 2013). All analyses were conducted with Mplus 6.1 (Muthén, & Muthén, 2010), using the robust Maximum Likelihood estimator (MLR) and 2000 random starts, 100 iterations for these random starts and the 100 bests retained for final stage optimization (Hipp & Bauer, 2006; McLachlan & Peel, 2000). All of the reported models converged on a replicated solution and can confidently be assumed to reflect a "real" maximum likelihood. For each parameterization (Figure 1, also see the Appendix), models with 1 to 12 latent profiles were estimated with the indicators' (SEEQ factor scores) intercepts and residuals freely estimated in all classes (Morin, Maïano, et al., 2011). Course level was controlled for in all analyses. For parsimony, only results from models with 1-8 classes are reported.

An important challenge in mixture modeling is determining the number of latent profiles in the data. Two important criteria used in this decision are the substantive meaning and theoretical conformity of the extracted profiles (Marsh, Lüdtke et al., 2009; Muthén, 2003) as well as the statistical adequacy of the solution (e.g., absence of negative variance estimates; Bauer & Curran, 2004). A number of statistical tests and indices are available to help in this decision process (McLachlan & Peel, 2000). Recent simulation studies indicate that four of these various tests and indicators are particularly effective in choosing the model which best recovers the sample's true parameters in mixture models (Henson, Reise, & Kim, 2007; McLachlan & Peel, 2000; Nylund, Asparouhov, & Muthén, 2007; Tofighi & Enders, 2008; Tolvanen, 2007; Yang, 2006): (i) the Consistent Akaïke Information Criterion (CAIC), (ii) the Bayesian Information Criterion (BIC), (iii) the sample-size Adjusted BIC (ABIC), and (iv) the Bootstrap Likelihood Ratio Test (BLRT). Additional simulation studies indicate that the ABIC and the classical Akaïke Information Criterion (AIC) are also effective in comparing models relying on different within-class specification, invariance assumptions, or parameterizations in line with those contrasted here (Lubke & Neale, 2006, 2008). In line with these results, these indicators (AIC, CAIC, BIC, ABIC, BLRT) will be reported. A lower value on the AIC, CAIC, BIC and ABIC suggests a better-fitting model. The BLRT is a parametric likelihood ratio test obtained through resampling methods that compares a *k*-class model with a *k-1*-class model. A significant *p* value indicates that the *k-1*-class model should be rejected in favor of a *k*-class model. Those studies also show that, when the indicators fail to retain the optimal model, the ABIC and BLRT tend to overestimate the number of classes, whereas the BIC and CAIC tend to underestimate it. Finally, the entropy indicates the precision with which the cases are classified into the various profiles. Although the entropy should not be used to determine the optimal number of profiles (Lubke & Muthén, 2007), it provides a useful summary of the classification accuracy. The entropy varies from 0 to 1, with higher values indicating less classification errors.

As emphasized by Marsh, Lüdtke et al. (2009) these tests are all variations of tests of statistical significance such that the so-called 'best' number of groups is heavily influenced by sample size. Although it might be reasonable to limit the number of groups when sample sizes are modest to avoid capitalizing on chance and enhancing replicability, it means that there is typically not an inherently correct number of groups. This is particularly relevant in the present investigation where the sample sizes is very large. Thus, as a complement, some (Morin, Maïano, et al., 2011; Petras & Masyn, 2010) suggest graphically presenting information criteria through "elbow plots" illustrating the gains associated with additional profiles. In these plots, the point after which the slope flattens out indicates the optimal number of profiles in the data. Interestingly, this approach relies more heavily on notions of variance explained that is less influenced by sample size than other approaches typically used. We note however that the efficacy of this strategy in helping to recover true population values has never been formally investigated in the context of simulation studies.

Following from Marsh and Bailey (1993), an additional index was computed for model including one to seven classes (selected from the elbow plots as realistic models) to reflect the stability of the estimated profiles within teachers among the longitudinal subsample of 195 teachers. For these models, we saved the posterior probabilities of membership in each profile associated with each set of class ratings into an external data file and conducted a repeated measure MANOVA on these class probabilities with teacher as the grouping variable and main effect. Contrasting with traditional methods of assigning individuals to a single profile by modal posterior probabilities, the present method avoids the biases associated with the categorization of continuous variables (MacCallum, Zhang, Preacher, & Rucker, 2002) and provides of more realistic representation of the data by considering the degree of likelihood of membership of the teachers in each profile (Marsh, Lüdtke et al., 2009). From these analyses, the eta^2 effect size measure as a reflection of the percentage of variance explained in the results by the factor was computed for each specific probability of class

membership, as well as for the full set of class probabilities, as a reflection of the stability of the estimated profiles (i.e., the variances in the class probabilities that can be attributed to teacher identities). The eta^2 was chosen for comparability with Marsh & Bailey (1993) results. Importantly, we propose this indicator as a useful summary of the stability of profile membership over time, and not as a criterion to be used in selecting the optimal representation of the data. Recommendations regarding the use of this indicator for any purpose other than purely descriptive would need to be guided by simulation studies. In the present context, this indicator is particularly interesting as we know from Marsh and Bailey's (1993) study that 47% of the variability of the ratings is due to the presence of stable *shape* effects, so that the final retained model is expected to result in a similar estimate.

Results

Comparisons of Alternative Models

We first examine the results from the four alternative models in terms of fit, classification accuracy, and stability.

Fit. The fit indices for the 1 to 8 profiles solutions across the four alternative parameterizations are reported in Table 1. When the recommended AIC and ABIC from models with similar numbers of profiles are compared, the results clearly show that Model 3 parameterization (i.e., factor mixture model) is superior to the various alternatives, but closely followed by Model 1 parameterization (i.e., the classical latent profile model), then Model 4 (i.e., including the higher-order effectiveness factor as control), with the worst results being associated with Model 2 (i.e., including the higher-order effectiveness factor as an profile indicator). The BIC and CAIC indicators support this conclusion. Based on these results and the fact that Model 3 is the one most in line with theoretical expectations a purely substantive investigation would likely retain Model 3 parameterization as the one providing the optimal representation of the data and simply ignore the results from the alternative parameterizations. Here, in line with our methodological objectives of illustrating these alternative models, we juxtapose the results from these four models, but reinforce that Model 3 should be retained on the basis of both theoretical and empirical criteria as providing the best representation of the data.

Classification accuracy. The entropy shows that the classification accuracy tends to be lower in models where *levels* of overall teacher effectiveness are partialled out in some way (Models 3 and 4) than in models were they are simply ignored (Model 1) or included in the mixture algorithm (Model 2). Thus, it seems that extracting naturally occurring level effects (i.e., reflecting the fact that teachers do differ from one another on their overall level of effectiveness rather than simply at the dimensional level), limit the classification accuracy of the models. This suggests that taking into account global *levels* of effectiveness may help in obtaining a better, more accurate, classification of teachers. This result reinforces the need to rely on strong theoretical bases and objective criteria in selecting the optimal model to best represent the data. However, in the present context, this result is not surprising. Indeed, from Marsh and Bailey (1993), we expect *level* effects to explain almost as much variance (37%) than *shape* (47%) effects in SEEQ ratings, so that disentangling both components should logically make it harder to classify teachers. This result also confirms the need to consider the global effectiveness factor as substantially meaningful in its own right, rather than as a simple artifact to be controlled for in the analyses. In fact, the remaining results suggest that, at least in this application, this loss in classification accuracy in well compensated by gains in terms of classification stability, and interpretability of the profiles as representing meaningful patterns of strengths and weaknesses.

Stability. Conversely, the stability of the estimated latent profiles within each teachers for the subsample of 195 teachers for whom repeated measures are available tend to be higher in the models were the overall teacher effectiveness *levels* are partialled out (Models 3 and 4) than in the other models (Models 1 and 2). Thus, although including *level* information appears to help the classification of teachers into profiles, this "improved" classification is in fact less useful due to its lower level of stability over time. Indeed, the results show that in order to estimate more stable profiles of teacher's

effectiveness – representing more useful guides for feedback interventions – *level* information should be extracted. In other words, *level* effects seem less stable than *shape* effects, confirming Marsh and Bailey (1993) results. Furthermore, looking closely at Model 1 results it is apparent that, although the average stability for the full set of profiles (.410 to .533) is close to the results from Model 3 (.434 to .570), the stability of each specific profile is more variable in Model 1, with some profiles presenting stability indices as low as .064 (i.e., only 6.4% of the variance in the probability of membership in these profiles can be attributed to stable teacher effects, versus at least 30.7% for Model 3). In other words, the results from Model 3 are clearly most in line with Marsh and Bailey's (1993) results.

Selecting the Optimal Number of Latent Profiles in the Solution

Further examination of the results reported in Table 1 in order to select the optimal number of latent profiles to retain reveals that the information criteria continue improving when latent profiles are added for each of the alternative parameterizations. This is not surprising given the large sample size and sample size dependency of these indicators. Indeed, for real data based on a large-enough sample size, the information criteria will always choose the most complex and, ultimately, the saturated model, as is apparently the case in the present investigation (Marsh, Lüdtke et al., 2009; Morin, Maïano, et al., 2011). Therefore, it has been recommended to complement this information with a theoretically grounded subjective evaluation of models including different number of classes, as well as on an inspection of parameters for statistical conformity (Marsh, Lüdtke et al., 2009; Marsh, Balla, & McDonald, 1988; Marsh, Hau, & Grayson, 2005; Muthén, 2003). We remind the reader here that, where this study purely substantive, the class enumeration process would be limited to Model 3.

In the present study, potentially due to the large sample size, all models were fully proper statistically. Similarly, multiple alternative solutions were fully interpretable and consistent with our a priori expectations regarding the nature of the profiles – particularly for the best fitting Model 3 – and converged on similar conclusions regarding the relative efficacy of the various models at extracting *level* effects from the profiles. Thus, consistent with previous recommendations (e.g., Morin, Maïano et al., 2011; Petras & Masyn, 2010) we relied on elbow plots to help in the selection of the final solution. The elbow plots for all four models are reported in the Appendix and generally converged on a five profile solution. More precisely, they seem to support a 4-5 profiles solution for Model 1, a 4-5 profiles solution for Model 2, a 5 profiles solution for Model 3, and a 5-6 profiles solution for Model 4. For the best fitting Model 3, the five and six profiles solutions were closely inspected and the six profiles solution did not add much to the interpretation of the results (i.e., splitting 1 class into two showing mostly *level* differences), confirming our reading of the elbow plot.

Interpretation of the Final Five Profiles Solutions

Given the methodological focus of the present study, the results from the five profiles solutions for all four models are reported in Figures 2 to 5 for comparison purposes. However, substantively, only Model 3 should be interpreted as providing the most appropriate representation of the data.

Model 1. Model 1 results (Figure 2) reveal clear *level* effects, showing that three of the latent profiles only differ from one another according to average *levels* of teacher effectiveness. The only exception is related to the distinction between the latent profiles 3 and 4. These profiles differ from one another on the Group Interaction and Individual Rapport dimensions (higher in profile 4) as well as on the Workload/Difficulty, Exams/Grading, and Assignments/Readings dimensions (higher in profile 3). Thus, these profiles correspond to the a priori differentiation we proposed between teachers that are mostly centered on affective relations with the students versus those that are mostly centered on cognitive objectives – a distinction that will be even more noteworthy in Model 3.

Model 2. As expected, the results from Model 2 (Figure 3) show even more pronounced *level* effects and present profiles differing from one another strictly according to teachers levels of global effectiveness. This observation and its convergence with our expectations, clearly argues against the usefulness of this parameterization in the context of the present study, but also more generally as a

way to maximize *level* effects in mixture models (which we argue is typically undesirable).

Model 3. The results from the retained factor mixture model (Model 3) are reported in Figure 4. As we anticipated, *level* effects remain apparent in some of these profiles (in profiles 1 and 4), reflecting the fact that some teachers are in fact generally "good" (profile 4, 24.7% of the sample) or "poor" (profile 1, 11.0% of the sample) across all of the evaluated dimensions of effectiveness. Forcing the extraction of these "residual" level effects (as in Model 4) would likely make no sense in the present application and create the false impression that all teachers present specific areas of required improvement or that all teachers only have specific strengths. Thus, although the "poor" teachers present average levels on the Workload/Difficulty dimension, they present very low levels on all remaining dimensions. This observation clearly indicates that it would be hard to prioritize specific areas of improvement for them, and that this prioritization should probably be done on an individual basis. Similarly, although the "good" teachers are generally less strong in the Group Interaction and Workload/Difficulty dimensions, saying that these reflect areas of improvement for them would neglect their generally high level of effectiveness on all dimensions. For these also, specific areas of improvement should probably be targeted on an individual basis.

However, the remaining three profiles differ more clearly according to their specific shape. Thus, profile 2 (25.1% of the sample) regroups generally average teachers, whose levels of Organization/Clarity, Workload/Difficulty, Exams/Grading and Readings/Assignments are generally satisfactory but who would do well to improve their levels of relational skills in the classroom (Enthusiasm, Group Interaction, Individual Rapport) as well as the Breadth of Coverage of the subject matter, and thus the overall Learning Value of the course. Going back to our theoretical expectations, this profile includes teachers that are clearly not oriented toward affective/relational objectives. Rather, they apparently mostly focus on performance goals in a purely cognitive perspective, ensuring adequate clarity, workload, evaluations and readings, but not going to extra mile to ensure mastery of the global subject area covered in the course. Then, profile 3 (20.3% of the sample) regroups more affectively/relationally oriented teachers presenting important strengths on the Enthusiasm, Group Interaction, and Individual Rapport dimensions. However, these strengths seem to occur at the detriment of sufficient Workload/Difficulty and Assignments/Readings, thus again impacting negatively the overall Learning/Value of the course. These teachers thus seem to focus mainly on performance goals in the classroom. Finally, profile 5 (18.8% of the sample) regroups generally good teachers, at least in terms of Enthusiasm, Organization/Clarity, Workload/Difficulty, Breadth of Coverage and most importantly, Learning Value. These teachers' main areas of improvement are related to Group Interaction and Individual Rapport, as well as to Exams/Grading. These teachers thus appears to clearly focus on cognitive (versus relational/affective) and mastery goals in the classroom.

Interestingly, the average stability of the profiles within the subsample of 195 teachers with repeated measures estimated from this solution is .507, meaning that 50.7% of the variability in the estimated probability of membership into the different profiles can be attributed to the teacher being evaluated rather than to situational variability. This very is very close to the 47% of the total variance in multidimensional SETs reported by Marsh and Bailey (1993) as attributable to teacher-specific stable *shape* effects (i.e., profiles) and higher than the 40.8% of total variance found to be attributable to profiles when we replicated Marsh & Bailey analyses on this larger data set; suggesting the greater precision of the present analysis in identifying teachers' profiles. Additional examination of the results showed that 50% of the teachers presented a clear dominant profile over time. For the others, 22% still did present a single dominant profile, while showing a higher level of fluctuations over time and 20% had two dominant profiles. In fact, only 8% of the teachers apparently presented unstable pattern of membership into the different profiles. A similar stability indicator was also computed for the higher-order effectiveness factor estimated in this solution. The within-teacher stability of this overall *level* of effectiveness is 36.2%, also quite close to the 37.1% reported by Marsh and Bailey (1993) as

attributable to teacher-specific stable *level* effects. These results confirm that the stability of teaching effectiveness profiles of is higher than the stability of teachers' overall levels of effectiveness.

Model 4. Figure 5 presents the results from Model 4, where global levels of effectiveness were directly partialled out from the nine SEEQ factors with a regression-control method. As expected, these results reflect pure *shape* effects, akin to the estimation of mean-centered profiles (where the overall *level* of effectiveness would have been subtracted from the profiles). A careful examination of these results supports the conclusions from the information criteria in showing that when compared with the factor mixture model (Model 3), this model provides a suboptimal representation of the data, both heuristically and statistically. It should be noted that the first and third profiles from Model 4 are similar to the fifth profile from Model 3 (cognitive-mastery orientation) and differ from one another on the relative strength of the Assignments/Readings, Workload/Difficulty, and Learning/Value, and factors (higher in the third profile). Similarly, the fourth profile from Model 4 is quite similar to the third profile (affective/relational-performance orientation) from Model 3. The second profile from Model 4 apparently corresponds to the first profile from Model 3 and regroups generally poor teachers that are otherwise good at Organization/Clarity, Exams/Grading and Assignments/Readings. These results thus suggest that the "poor" teachers identified in Model 3 should probably not focus on these more technical targets if they aim to improve their overall teaching effectiveness. Finally, the remaining fifth profile from Model 4 apparently has no direct correspondence in Model 3 and regroups generally average teachers that would do well to focus on improving the Learning/Value, Enthusiasm and Workload/Difficulty of their teaching. The results obtained in the present study under this parameterization are harder to interpret than those from the factor mixture models and provide a worse representation of the data according to the fit indices. However, the observed pattern of results suggest that this model may still represent an efficient method of partialling out *level* effects and thus, could likely represent a viable alternative to factor mixture models when they fail to sufficiently partial out *level* effects. Also, as shown here, comparisons of these models may also help to provide alternative perspectives on the results, especially in relation to specific profiles which remain mostly defined by levels differences in the factor mixture operationalization (Model 3).

Discussion

This study is a methodological-substantive synergy aimed at contrasting alternative methods of partialling out *level*-related quantitative differences from the profiles when such effects are present in order to maximize *shape*-related qualitative differences between the profiles and thus to increase their theoretical meaningfulness and practical utility. Simultaneously, in order to illustrate these methods, this study aimed at identifying profiles of teachers based on multidimensional students' ratings of their effectiveness. In the present case, level effects were related to generic teacher effectiveness, indicative of the fact that teachers are more or less good or bad generally notwithstanding their specific profiles of strengths and weaknesses. However, the models tested here and the conclusions have broad relevance to any person-centered investigation of multidimensional construct where level effects are present and explain part of the correlations between the various dimensions underlying the construct of interest and thus are be particularly relevant to the field of performance appraisal.

Methodological Implications: Disentangling Shape and Level in Latent Profiles Models

It is generally recognized that construct validation procedures are important to determine whether the extracted latent profiles can really be interpreted as representing meaningful subgroups of participants (e.g., Bauer & Curran, 2004; Muthén, 2003). However, one implicit criterion that permeates the person-centered literature (e.g., Bauer, 2007, De Boek et al., 2005) is the need to observe *shape*-related qualitative differences between the extracted profiles. Without clear shape differences, then the main assumption is that the data would be best represented by continuous latent factors, rather than categorical latent profiles. However, for some multidimensional constructs, both level and shape effects can be strong, making the identification of qualitatively different profiles

harder since strong *level* effects may create equally strong quantitative differences between the profiles (e.g. Masyn et al., 2010). Indeed, classical LPA assume the conditional independence of the indicators, meaning that conditional on class membership, the residual correlations between the various dimensions are assumed to be zero. Thus, when generic quantitative *level* effects are present, such as it is the case when evaluating multiple interrelated competencies, these effects create conditional dependencies that have no other choice than to be absorbed by the latent profiles when they are not specifically modeled. Here, we contrasted four alternative models in order to find a way to extract unnecessary *level* effects from a LPA solution: (a) a classical LPA model assuming the conditional independence of the dimensions (Model 1); (b) a classical LPA model assuming the conditional independence of the dimensions but including a higher-order generic factor as an additional indicator (Model 2); (c) A factor mixture models including a generic continuous factor to account for *level*-based conditional dependencies between the indicators (Model 3); (d) a LPA model in which a higher-order generic factor was added as a control variable on which the main dimensions were regressed prior to the estimation of the main LPA model (Model 4). We also argued that these four models relied on highly different substantive assumptions: (a) Model 1 assumes that *level* effects would be negligible; (b) Model 2 assumes that *level* effects would be present, but that they simply should be considered as an additive indicator of the profiles; (c) Model 3 assumes that both *shape* and *level* effects would be present, meaningful in themselves, and complementary in nature; (d) Model 4 also assumes that both *shape* and *level* effects would be present, but also that level effects simply represent a biasing influence that needs to be controlled for a clearer investigation of shape-based profiles. Thus, an important difference between Models 3 and 4 is that Model 3 assumes that the continuous latent factor needs to be considered as substantively meaningful in its own right, potentially as meaningful as the latent profiles themselves, whereas Model 4 assumes that this global dimension is simply some form of bias to be extracted from the latent profiles.

In the present application, the results clearly showed, as expected, that Model 2 was inappropriate and amplified the quantitative *level* differences between profiles. Model 1 also confirmed our expectations formed on the basis of Marsh and Bailey (1993) results, revealing the presence of strong shape and level effects in the extracted latent profiles. Model 3 provided the clearest results according to substantive interpretations and also provided a better representation of the data according to the information criteria considered. Conversely, Model 4 provided latent profiles that showed even purer qualitative *shape* differences and provide an interesting complement of information to the profiles identified based on Model 3. However, this model did provide a worse fit to the data according to the information criteria considered, was harder to interpret in the present study, and did not meet our theoretical and empirical expectations that rather supported Model 3. The main differences between the solutions obtained with Models 3 and 4 are that Model 4 forced all *level*-related information out of the estimated latent profiles whereas Model 3 estimated both the *level* (common factor) and *shape* (profiles) effects simultaneously and thus ended up extracting only unnecessary, or residual, *level* effects from the estimated profiles. Model 3 thus apparently provides a more organic representation of the data when the extraction of meaningful latent profiles requires at least some *level*-based distinctions, such as in the present case where latent were needed to reflect the fact that some teachers are simply globally *good* or *bad* across all dimensions of SETs considered. In such cases, forcing the extraction of all *level* effects (Model 4) will likely result in less meaningful profiles. However, in some cases where *level* effects are smaller, unnecessary, or seen as a potentially biasing influence, Model 4 may be more appropriate. For the present study, Marsh and Bailey (1993) results clearly showed that although *shape* effects were more pronounced than *level* effects, both *shape* and *level* effects were important in SETs.

We thus recommend that future research aiming at disentangling *shape* from *level* effects on latent profiles analyses should rely on strong theoretical and/or empirical a priori expectations

regarding the presence, nature, and meaning of both *shape* and *level* effects in the data. From these expectations, a choice should be made, a priori, between models 2-3-4 in order to pick the model best suited to these specific expectations. Then, the retained model should be empirically contrasted with Model 1 on the basis of information criteria in order to directly investigate the added value of bringing *level* effects in the latent profile model. Realistically, Model 1 (a classical LPA) could first be estimated and its solution examined for indications of strong *level* effects needing to be taken into account and consistent with theoretical expectations. To this end, Marsh and Bailey (1993) (M)ANOVA based procedure can help to form clearer expectations when applicable to the data set under consideration (i.e., including repeated multidimensional assessments). When strong level effects are present and detract from the meaningfulness of the profiles, while theoretical expectations strongly suggest that shape effects should also be present in the data, then some form of conditional dependence needs to be accounted for in the models. Although this can be done in multiple manners, we proposed Models 2, 3 and 4 parameterisations for cases were the dimensions are assumed to also form a single higher-order dimension. Here this construct was assumed to reflect global teaching effectiveness, but a similar cases can be built for other constructs such as global commitment (e.g., Morin, Morizot et al., 2011), or even global self-concept (e.g., Marsh, Lütdke et al., 2009). Following Lubke & Neale (2006, 2008), these two alternative parameterisations (Model 1 versus the model retained as best suited to the specific investigation) can then be contrasted both substantively and in terms of fit so as to retain the most appropriate representation of the data. Although Model 3 proved best in the present investigation, we do not believe that this conclusion can be generalized to all research contexts. Interestingly, Model 4 also provide a way to account for rater biases and shared method variance, which can represent important issues in the assessment of job competencies (Latham, & Mann, 2006; Fletcher, 2001). Although this was not a major issue in the present study where teachers effectiveness ratings were already based on class-average ratings from multiple students, the common higher-order effectiveness factor controlled for in Model 3 would also absorbs shared method variance, in addition to meaningful variance related global effectiveness (Eid et al., 2008; Podsakoff, MacKenzie, & Podsakoff, 2012) – an issue that should be kept in mind when appropriate. These models can easily be extended to include multiple method factors for different types of raters, as well as a global effectiveness factor.

Looking at the classification accuracy of the various models and at the within-teacher stability in probabilities of class membership proved also highly informative in showing that the extraction of *level* effects results in somewhat poorer classification accuracy (i.e., teachers were harder to classify in the various profiles) but in a greater level of stability in the resulting classification. Interestingly, for the retained five class solution, Model 3 provided the greatest level of classification stability. These results suggest that *levels* of generic effectiveness provide valuable information in performance appraisals. However, excluding this generic *level* from the assessment of the profiles apparently helps to identify more stable core mechanisms underlying teaching style. For intervention purposes, we argue that targeting these core mechanisms is likely to be more efficient than simply targeting overall level of effectiveness (e.g. Marsh, 2007b; Marsh & Roche, 1993).

An important methodological issue that would need to be considered in the context of future studies has to do with the best way to contrast, and compare, these different models. We have strongly argued that clear substantive and empirical a priori should be used to guide the selection, and evaluation, of the model assumed to be bested suited to the investigation. Then, we propose that the retained Model (2-3-3) should be contrasted with Model 1 on the basis of commonly used information criteria, in order to directly test the assumption that there are indeed level effects present in the data and that these need to be systematically considered. However, previous investigations regarding the efficacy of the various information criteria in choosing between differently parameterized models including the same number of classes, show that these work, but that their efficacy remains limited and

may change under different conditions (e.g. Lubke & Neale, 2006, 2008). Clearly, additional investigations are needed in this area. For the meantime, as in most applications of mixture modeling, some part of the decision process must remain subjective, a main limitation of these methods underlying the need to ground such decisions in clear theoretical bases. However, as emphasized by Marsh et al. (1988, 2005), Hu and Bentler (1998, 1999), and others (e.g., Bentler & Bonett, 1980; Browne & Cudeck, 1993; Cudeck & Henly, 2003; Jöreskog & Sörbom, 1993; Muthén, 2003) data interpretations and their defence ultimately remains to some extent a subjective undertaking that requires researchers to immerse themselves in their data.

Similarly, although previous studies have shown that distributional tests and classification accuracy (i.e., entropy) should not be used to select the optimal number of latent classes present in the data (e.g., Henson et al., 2007; Tofighi & Enders, 2008), investigations of their efficacy in contrasting alternative models including the same number of latent profiles are more limited (Lubke & Neale, 2008). This issues should clearly be investigated in the context of future studies. Similarly, alternative distributional indices of class separation (e.g. Mahalnobis distance) or factor scores distributions (see, for instance, Steinley & McDonald, 2007) should also be investigated in the context of studies were the population generating model is know – which was not the case in the present study. Finally, although we presented an additional MANOVA-based eta^2 indicator that could be used to describe the stability of the estimated profiles within the nested longitudinal subsample of teachers, we must stress that the generalizability of this indicator would be limited to studies including such longitudinal data. Similarly, the use of this indicator, pending systematic investigation of its efficacy in the context of simulation studies, should at this stage remain purely descriptive.

Substantively, the current results raised a number of interesting issues and directions for future research that we will discuss in the next section, but some limitations must also be taken into account. Indeed, an important criterion against which to evaluate the meaningfulness of person-centered solutions, or any other statistical result for that matter, has to do with their generalizability to new samples. Without replication, any substantive result remains tentative. Here, the fact that the extracted profiles proved to be mostly in line with our theoretical expectations and could be replicated over time within the longitudinal subsample gives additional credibility to the conclusions. However, there are potential biases associated with the fact that the sample comes from a single US University and thus that the longitudinal sample includes teachers who had worked at this university for an extended period. However, given that tenure decisions are mostly based in research track record in research-oriented universities, this potential bias probably had little effect on the longitudinal component of the present investigation. Indeed, the results based on this longitudinal subsample are largely consistent with results based on cross-sectional research from the same university (e.g., Marsh & Roche, 2000) that does not suffer from the same selection bias. Unfortunately, there is a dearth of other longitudinal studies with which to evaluate the generalizability of these findings (Marsh, 2007b). Regarding the full sample, the question of whether the results can be expected to generalize to other universities remains open. However, results based on multiple published studies based on this sample show good generalizability to results from research literature on teaching effectiveness (see reviews by Marsh, 1987, 2007b), giving some credence to the present results. However, pending replication, their generalizability to other samples, universities, and countries is a question that should be systematically investigated in the future. Although, this generalizability is not so much of a concern in terms of the methodological implications that are the core of this study, this limitation must be kept in mind as we now move to discussions of substantive interpretations of our results and likely practical implications.

Substantive Implications: Profiles of University Teacher Effectiveness.

Substantively, this study attempted to build a taxonomy of University teachers according to multidimensional ratings of their effectiveness by students. Although a few previous studies attempted to profile teachers according to their effectiveness (e.g., Brekelmans et al., 1993; Rickards et al., 2005;

Sang et al., 2009; Winterbottom et al., 2008), this is the first study to specifically target university teaching, to rely on an extensive multidimensional conceptualization of SETs, and to explicitly extract quantitative level differences from the profiles in order to obtain cleaner and more meaningful qualitative shape differences between these profiles. The results from the final retained model revealed the presence of five latent profiles of teachers. Membership into these profiles was generally quite stable over time, more so than global ratings of teaching effectiveness, indicating that overall appreciations of a teacher by students may fluctuate more from one class to the other than specific ratings regarding strength and weaknesses of the assessed teachers. However, this stability was not perfect. Indeed, teacher identify only accounted for 50% of the membership into these specific profiles, suggesting that teachers can indeed improve, or change, over time. This result is encouraging for feedback interventions and shows that even at the personal profile level, the pattern of strengths and weaknesses of teachers can change. However, this result also suggest that before assigning teachers to specific profiles, they should be assessed in various contexts (different levels and subjects) over at least two years in order to obtain a precise idiographic picture of their individual profiles.

Interestingly, one of the largest profiles (25%) included generally good teachers across all dimensions of teaching effectiveness. For these teachers, no specific area of weakness could be identified. Similarly, the smaller (11%) subgroup of "poor" teachers had low results across all dimensions of SETs. However, when this same subgroup was examined under Model 4 parameterization, a more refined picture emerged. Model 4 results suggests that once their overall levels of effectiveness is extracted, these "poor" teachers remain relatively good at Exams/Grading, Organization/Clarity, and Assignments/Readings and should not focus on these more technical targets in attempts to improve their teaching effectiveness.

One large (25%) profile seems to be particularly in need of feedback intervention as their general level of effectiveness is quite average. These teachers present levels of Organization/Clarity, Workload/Difficulty, Exams/Grading and Readings/Assignments that are satisfactory. However, their effectiveness in less technical areas of teaching appear to be in need of improvement. Indeed, they seem to be lacking both at the level of their relational skills in the classroom (Enthusiasm, Group Interaction, Individual Rapport) and regarding Breadth of Coverage and Learning/Value. It seems like they are teaching the basics, without getting personally involved or investing too much energy in their teaching activities. These teachers apparently mainly focus on performance goals in a purely cognitive perspective, ensuring adequate clarity, workload, evaluations and readings, but not going to extra mile to ensure extensive mastery of the subject area. These teachers, together with the "poor" teachers, would probably benefit from training programs encompassing relational skills and breadth of coverage. Future studies contrasting different ordering of these training components (relational skills versus coverage) to verify whether intervening on one component could have rippling effects on the other would likely be very informative.

The last two profiles are even more convergent with our theoretical expectations and each regroups close to 20% of the sample. One of those includes affectively/relationally oriented teachers focusing mostly on performance goals in the classroom. These teachers have important strengths on the Enthusiasm, Group Interaction, and Individual Rapport dimensions, which occur at the detriment of sufficient Workload/Difficulty and Assignments/Readings, thus negatively influencing the Learning/Value of the course. Training programs for these teachers should target the improvement of course learning value through increased Workload/Difficulty and Readings/Assignments, addressing strategies to ensure that these augmentations do not occur at the detriment of maintaining quality relationships with students. Conversely, the last profile regroups teachers that are generally good, enthusiastic, clear and organized, that tend to use appropriate levels of Workload/Difficulty, and whose courses include a sufficient Breadth of Coverage to ensure elevated Learning Value for students. These teachers appear to focus mainly on cognitive mastery goals in the classroom and, as

such, their main area of improvement is related to Group Interaction and Individual Rapport, as well as to Exams/Grading. This last result regarding Exams/Grading could potentially be explained in two alternative manners that should be more systematically contrasted and investigated in future studies. On the one hand, the mastery focus of these teachers may lead them to attribute less importance to formal evaluation and grading practices since evaluating/ranking students is more closely related to performance versus mastery goals. On the other hand, their broader coverage of the subject may lead them to develop more extensive evaluations procedures covering both central and peripheral facets, leading to a lower level of satisfaction among students who may not be used to such broad evaluations procedures. For these teachers, training programs should likely target the improvement of relational facets of teaching and the development of evaluations procedures more closely related to the objective of the program. It would be interesting, when possible, to use dyadic programs where teachers from the "cognitive-mastery" profile would work with teachers from the "relational-performance" profile in order to share the specific teaching strategies that are so successful for each of these profiles.

References

Allport, G. (1937). *Personality: A psychological interpretation*. New York: Holt, Rinehart & Winston.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261-271.

Bailey, K. D. (1994). *Typologies and taxonomies: An introduction to classification techniques*. Thousand Oaks, CA: Sage.

Bauer, D.J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research, 42*, 757-786.

Bauer, D.J. & Curran, P.J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 9*, 3-29.

Bauer, D.J. & Shanahan, M.J. (2007). Modeling complex interactions: Person-centered and variable-centered approaches. Little, T.D., Bovaird, J.A. & Card, N.A. (Eds.). *Modeling ecological and contextual effects in longitudinal studies of human development* (pp. 255-283). Mahwah, NJ: LEA.

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

Bergman, L.R., Magnusson, D., & El-Khouri, B.M. (2003). *Studying individual development in an interindividual context: A person-oriented approach*. New York, NY: Erlbaum.

Bergman, L.R., & Trost, K. (2006). The person-oriented Vs. the variable-oriented approach: Are they complementary, opposites, or different worlds? *Merrill-Palmer Quarterly, 52*, 601–632.

Black, P., McCormick, R., James, M., & P edder, D. (2006). Assessment for learning and learning how to learn: A theoretical enquiry. *Research Papers in Education, 21*, 119-132.

Brekelmans, M., Levy, J., & Rodriguez, R. (1993). A typology of teacher communication style. In T. Wubbels & J. Levy (Eds.), *Do you know what you look like?* (pp. 46–55). London, UK: Falmer.

Brewer, M.B. (1991). The social self: On being the same and different at the same time. *Personality & Social Psychology Bulletin, 17*, 475–482.

Brok, P.D., Fisher, D., Brekelmans, M., Wubbels, T., & Rickards, T. (2006). Secondary teachers' interpersonal behavior in Singapore, Brunei and Australia: A cross-national comparison. *Asia Pacific Journal of Education, 26*, 79–95.

Canrinus, E.T., Helms-Lorenz, M., Beijaard, D., Buitink, J., & Hoffman, A. (2011). Profiling teachers' sense of professional identity. *Educational Studies, 37*, 593-608.

Cattell, R.B. (1952). The three basic factor-analytic designs: Their interrelations and derivatives. *Psychological Bulletin, 49*, 499–520.

Chen, F.F., West, S.G., & Sousa, K.H. (2006). A comparison of bifactor and second-order models of quality of life. *Structural Equation Modeling, 41*, 189-225.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences, 2nd Ed. Hillsdale, NJ: Erlbaum.

Cross, S.E., & Madson, L. (1997). Models of the self: Self-construals and gender. P*sychological Bulletin, 122,* 5-37.

Cudeck, R., & Henly, S.J. (2003). A realistic perspective on pattern representation in growth data: Comment on Bauer and Curran (2003). *Psychological Methods, 8,* 378-383.

De Boek, P., Wilson, M., & Acton, G.S. (2005). A conceptual framework for distinguishing categories and dimensions. *Psychological Review, 112,* 129-158.

Eid, M., Nussbeck, F.W., Geiser, C., Cole, D.C., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: different models for different types of methods. *Psychological Methods, 13,* 230-253.

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.* Agathon, New York, pp. 368–395.

Fletcher, C. (2001). Performance appraisal and management: The developing research agenda. *Journal of Occupational and Organizational Psychology, 74,* 473-487.

Hativa, N. (1996). University instructors' ratings profiles: Stability over time, and disciplinary differences. *Research In Higher Education, 37,* 341–365.

Helgeson, V.S. (1994). Relation of agency and communion to well-being: Evidence and potential explanations. *Psychological Bulletin, 116,* 412-428.

Henson, J. M., Reise, S. P. & Kim, K. H. (2007).. Detecting mixtures from structural model differences using latent variable mixture modeling: a comparison of relative model fit statistics. *Structural Equation Modeling, 14,* 202-226.

Hershberger, S.L. (2006). The problem of equivalent structural models. In G.R. Hancock, & R.O. Mueller (Eds). Structural Equation Modeling, A second course (pp. 13-41). Greenwich, CT: Information Age.

Hipp, J.R., & Bauer, D.J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods, 11,* 36-53.

Hobson, C.J. & Gibson, F.W. (1983). Policy capturing as an approach to understanding and improving performance appraisal: A review of the literature. *Academy of Management Review, 8,* 640-649.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41-54.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

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 .

James, M., & Pedder, D. (2006). Beyond method: Assessment and learning practices and values. *The Curriculum Journal, 17,* 109-138.

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Lincolnwood, IL: Scientific Software International.

Krueger, R.F., Markon, K.E., Patrick, C.J., & Iacono, W.G. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implication for DSM-IV. *Journal of Abnormal Psychology, 114,* 537-550.

Kuo, P.-H., Aggen, S.H., Prescott, C.A., Kendler, K.S., & Neale, M.C. (2008). Using factor mixture modeling approach in alcohol dependence in a general population sample. *Drug and Alcohol Dependence, 98,* 105-114.

Latham, G.P., & Mann, S. (2006). Advances in the science of performance appraisal: Implications for practice. In G.P. Hodgkinson & J.K. Ford (Eds.), *International Review of Industrial and Organizational Psychology Vol. 26* (pp. 295-337). Hoboken, NJ: Wiley.

Lubke, G.H., & Muthén, B.O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10,* 21-39.

Lubke, G.H., & Muthén, B.O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling, 14*, 26-47.

Lubke, G. & Neale, M. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41*, 499-532

Lubke, G. & Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models? *Multivariate Behavioral Research, 43*, 592-620

Lubke, G. & Tueller, S. (2010). Latent class detection and class assignment: A comparison of the MAXEIG taxometric procedure and factor mixture modeling approaches. *Structural Equation Modeling, 17*, 605-628.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19-40.

Magidson, J., & Vermunt, J.K. (2002). Latent class models for clustering: A comparison with K-Means. *Canadian Journal of Marketing Research, 20*, 37-44.

Marsh, H.W. (1982). 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. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253-388. (Whole Issue No. 3)

Marsh, H.W. (2007a). *Self-concept theory, measurement and research into practice: The role of self-concept in Educational Psychology*. Leicester, UK: British Psychological Society.

Marsh, H.W. (2007b). Students' evaluations of university teaching: Dimensionality, reliability, validity, biases, and usefulness. In R.P. Perry & J.C. Smart (Eds.), *The scholarship of teaching and learning in higher education: Evidence-based perspective* (pp. 319–383). Dordrecht, NL: 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. & 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., & Hau, K.-T. (2007). Latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology, 32*, 151-171.

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*. Hillsdale, NJ: Erlbaum.

Marsh, H. W., & Hocevar, D. (1991). 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., Lüdtke, O., Trautwein, U., & Morin, A.J.S. (2009). Latent Profile Analysis of Academic Self-concept Dimensions: Synergy of Person- and Variable-centered Approaches to the Internal/External Frame of Reference Model. *Structural Equation Modeling, 16*, 1-35.

Marsh, H.W., Muthén, B.O., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A.J.S., & Trautwein, U. (2009). Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching. *Structural Equation Modeling, 16*, 439-476.

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. (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.

Masyn, K.E., Henderson, C.E., & Greenbaum, P.E. (2010). Exploring the latent structures of psychological constructs in social development using the dimensional-categorical spectrum. *Social Development, 19*, 473-493.

McClelland, D.C. (1987). *Human motivation.* Cambridge, NY: Cambridge University.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

Meece, J.L., Anderman, E.M., & Anderman, L.H. (2005). Classroom goal structure, student motivation and academic achievement. *Annual Review of Psychology, 57*, 487-503.

Midgley, C. (2002). *Goals, goal structures, and patterns of adaptive learning*. Mahwah, NJ: Erlbaum.

Morin, A.J.S., Maïano, C., Nagengast, B., Marsh, H.W., Morizot, J., & Janosz, M. (2011). General growth mixture analysis of adolescents' developmental trajectories of anxiety: The impact of untested invariance assumptions on interpretations. *Structural Equation Modeling, 18*, 613-648.

Morin, A.J.S., Marsh, H.W., & Nagengast, B. (2013). Exploratory Structural Equation Modeling. In G.R. Hancock & R.O. Mueller (Eds.), *Structural Equation Modeling: A Second Course, 2nd Edition* (pp. 395-436). Greenwich, Connecticut : IAP.

Morin, A.J.S., Morizot, J., Boudrias, J.-S., & Madore, I. (2011). A multifoci person-centered perspective on workplace affective commitment: A latent profile/factor mixture Analysis. *Organizational Research Methods, 14*, 58-90.

Morin, A.J.S., Tran, A., & Caci, H. (2013). Factorial Validity of the ADHD Adult Symptom Rating Scale in a French community sample: Results from the ChiP-ARD study. *Journal of Attention Disorders*. Online first: DOI: 10.1177/1087054713488825.

Morin, A.J.S., Rodriguez, D., Fallu, J.-S., Maïano, C., & Janosz, M. (2012). Academic Achievement and Adolescent Smoking: A General Growth Mixture Model. *Addiction.*

Muthén, B.O. (2001). Latent variable mixture modeling. In G.A. Marcoulides, & R.E. Schumaker (Eds.), *New developments in Structural Equation Modeling* (pp.1-33). Mahwah, NJ: Erlbaum.

Muthén, B.O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*, 81-117.

Muthén, B.O. (2003). Statistical and Substantive Checking in Growth Mixture Modeling: Comment on Bauer and Curran (2003). *Psychological Methods, 8*, 369-377.

Muthén, B.O. (2006). Should substance use disorders be considered as categorical of dimensional? *Addiction, Suppl 1*, 6-16.

Muthén, B.O. & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors, 31*, 1050-1066.

Muthén, B.O. & Asparouhov, T. (2009). *Growth mixture modeling: Analysis with non-Gaussian random effects*. In Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (eds.), Longitudinal Data Analysis, pp. 143-165. Boca Raton: Chapman & Hall/CRC Press.

Muthén, B.O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*, 463-469.

Muthén, L.K., & Muthén, B.O.(1998-2010). *Mplus user's guide.* Los Angeles CA: Muthén & Muthén.

Myers, I.B., & McCaulley, M.H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.

Nagin, D.S. (2010). Group-based trajectory modeling: An overview. In A.R. Piquero, & D. Weisburd (Eds.), *Handbook of Quantitative Criminology* (pp. 53-67). New York, NY: Springer.

Nylund, K.L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535-569.

Petras, H., & Masyn, K. (2010). General growth mixture analysis with antecedents and consequences of change. In A.R. Piquero, & D. Weisburd (Eds.), *Handbook of Quantitative Criminology* (pp. 69-100). New York, NY: Springer.

Oscarson, D.J., & Finch, C.R. (1979). *Adoption-proneness among trade and industrial teachers as*

measured by cluster analysis. Paper presented at the 73rd annual convention of the American Vocational Association, Anaheim, California.

Pedder, D. (2007). Profiling teachers' professional learning practices and values: differences between and within schools. *The Curriculum Journal, 18*, 231-252.

Podsakoff, P.M., MacKenzie, S.B., & Podsakoff, N.P. (2003). Sources of method bias in social science and recommendations on how to control it. *Annual Review of Psychology, 63*, 539-569.

Prawat, R.S. (1985). Affective versus cognitive goal orientations in elementary teachers. *American Educational Research Journal, 22*, 587-604.

Reb, J., & Greguras, G.J. (2010). Understanding performance ratings: Dynamic performance, attributions, and rating purpose. *Journal of Applied Psychology, 95*, 213-220.

Reise, S.P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696.

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.

Rickards, T., Brok P.D. & Fisher D. (2005). The Australian science teacher: a typology of teacher–student interpersonal behaviour in Australia. *Learning Environments Research, 8*, 267–287.

Roseth, C.J., Johnson, D.W., & Johnson, R.T. (2008). Promoting early adolescents' achievement and peer relationships: The effects of cooperative, competitive, and individualistic goal structures. *Psychological Bulletin, 134*, 223-246.

Sang, G., Valcke, M., Braak, J.V., & Tondeur, J. (2009). Investigating teachers' educational beliefs in Chinese primary schools: Socioeconomic and geographical perspectives. *Asia-Pacific Journal of Teacher Education, 37*, 363-377.

Shavelson, R.J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions and behaviors. *Review of Educational Research, 51*, 455-498.

Steinley, D., & Brusco, M.J. (2011). Evaluating mixture modeling for clustering: recommendations and cautions. *Psychological Methods, 16*, 63-79.

Steinley, D., & McDonald, R.P. (2007). Examining factor scores distributions to determine the nature of latent spaces. *Multivariate Behavioral Research, 42*, 133-156.

Sterba, S.K. & Pek, J. (2012). Individual influence on model selection. *Psychological Methods, 17*, 582-599.

Swank, P.R., Taylor, R.D., Brady, M.P., & Freiberg, H.J. (1989). Sensitivity of classroom observation systems: Measuring teacher effectiveness. *Journal of Experimental Education, 57*, 171-186.

Tay, L., Newman, D.A., & Vermunt, J.K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods, 14*, 147-176.

Tofighi, D., & Enders, C. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), Advances in latent variable mixture models (pp. 317-341). Charlotte, NC: Information Age Publishing.

Tolvanen, A. (2007). *Latent growth mixture modeling: A simulation study*. Doctoral dissertation, Department of Mathematics, University of Jyvaskyla, Finland.

Uebersax, J.S. (1999). Probit Latent Class Analysis with Dichotomous or Ordered Category Measures: Conditional Independence/Dependence. *Applied Psychological Measurement, 23*, 283-297.

Vermunt, J.K. (2011). K-Means may perform as well as mixture model clustering but may also be much worse: Comment on Steinley and Brusco (2011). *Psychological Methods, 16*, 82-88.

Vermunt, J.K. & Magidson, J. (2000). *Latent GOLD 1.0 User's Manual*. Boston: Statistical Innovations.

Vermunt, J.K., & Magidson, J. (2002). Latent class cluster analysis. In J. Hagenaars & A.

McCutcheon (Eds.), *Applied latent class models* (pp. 89-106). New York: Cambridge.

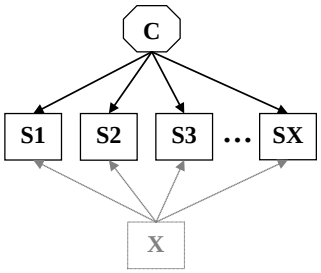
Waller, N.G., & Meehl, P.E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA: Sage.

Wang, C.K.J., & Liu, W.C. (2008). Teachers' motivation to teach national education in Singapore: A self-determination theory approach. *Asia Pacific Journal of Education, 28*, 395-410.

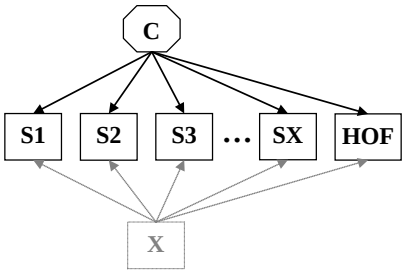
Winterbottom, M., Taber, K., Brindley, S., Fisher, L., Finney, J., & Riga, F. (2008). Understanding differences in teachers' values and practices in assessment. *Teacher Development, 12*, 15-35.

Woolley, S. L., Benjamin, W. J. J., & Woolley, A. W. (2004). Construct validity of a self-report measure of teacher beliefs related to constructivist and traditional approaches to teaching and learning. *Educational and Psychological Measurement, 64*, 319-331.

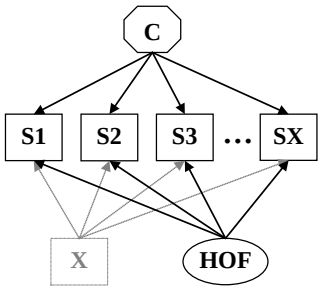
Yang, C. (2006). Evaluating latent class analyses in qualitative phenotype identification. *Computational Statistics & Data Analysis, 50*, 1090–1104.



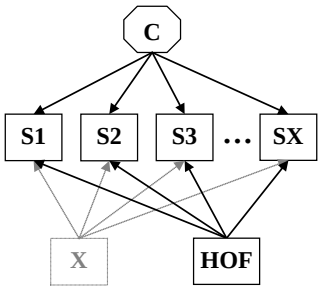
Model 1: Classical latent profile analysis.



Model 2: Latent profile analysis incorporating the higher order factor as class indicator.



Model 3: Factor mixture analysis with a class invariant higher order latent factor.



Model 4: Latent profile analysis incorporating the higher order factor as an additional control.

Figure 1. *Alternative models considered in the present study.*
Note. Squares represent observed variables; ovals represent continuous latent variables; octagons represent categorical latent variables; grayscale reflects the potential inclusion of controls variables; S1-SX represent the main scales of the multidimensional construct being assessed, or more generally the main mixture indicators; X represent the controlled variables; HOF represent the higher order continuous latent factor; C represent the categorical latent class.

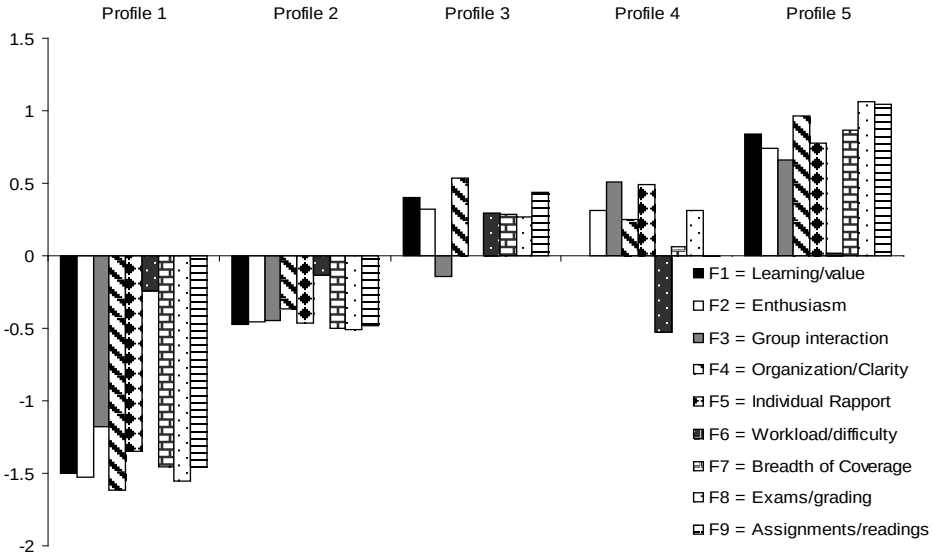


Figure 2. *Results from the latent profiles models based on 9 factors (Model 1)*

Figure 3. *Results from the latent profiles models based on 10 factors (Model 2)*

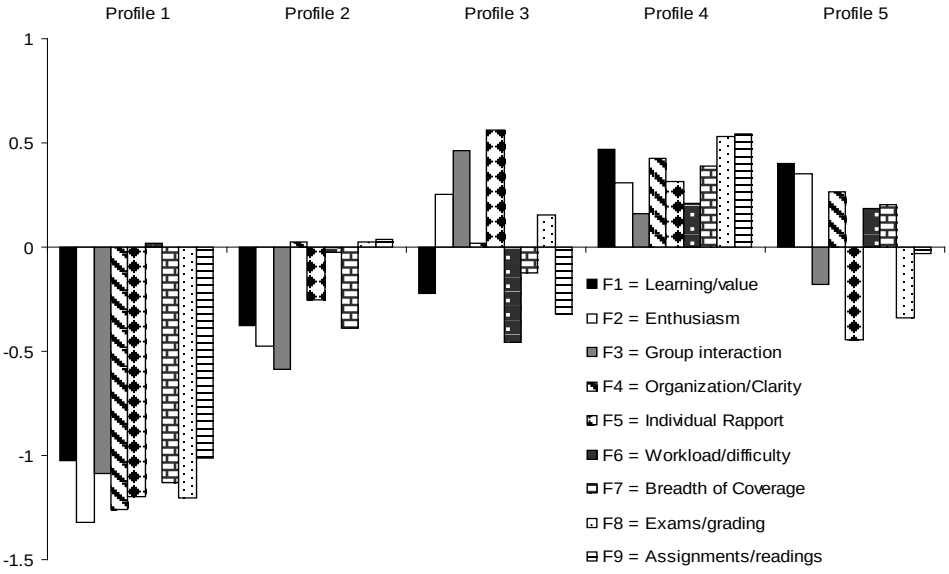


Figure 4. *Results from the factor mixture models based on 9 factors (Model 3)*

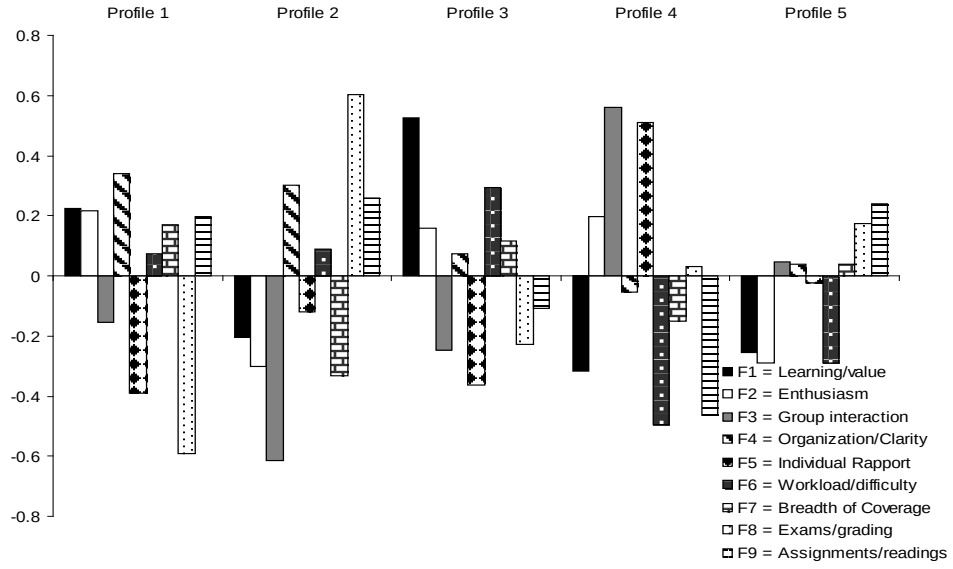


Figure 5. *Results from the models based on 9 factors controlling for the higher-order factor (Model 4).*

Table 1.
Fit Indices from Alternative Models 1 to 4.

Model	LL	#fp	Scaling	AIC	CAIC	BIC	ABIC	Entropy	BLRT	Stability M (Range)
Latent profiles models based on 9 factors (Model 1)										
1 Class	-398121	27	1.108	796297	796550	796523	796437	Na	Na	Na
2 Class	-361683	46	1.200	723458	723890	723844	723697	0.824	≤ 0.001	Na
3 Class	-351182	65	1.253	702494	703103	703038	702832	0.810	≤ 0.001	0.533 (0.393-0.658)
4 Class	-347316	84	1.248	694799	695586	695502	695235	0.797	≤ 0.001	0.330 (0.064-0.998)
5 Class	-345552	103	1.397	691310	692275	692172	691845	0.736	≤ 0.001	0.425 (0.262-0.586)
6 Class	-344038	122	1.790	688320	689463	689341	688953	0.722	≤ 0.001	0.414 (0.064-1.000)
7 Class	-342544	141	1.373	685370	686692	686551	686103	0.725	≤ 0.001	0.410 (0.064-1.000)
8 Class	-341637	160	1.415	683594	685093	684933	684425	0.702	≤ 0.001	Na
Latent profiles models based on 10 factors (Model 2)										
1 Class	-492609	30	1.114	985277	985558	985528	985433	Na	Na	Na
2 Class	-447738	51	1.238	895578	896056	896005	895843	0.858	≤ 0.001	Na
3 Class	-431111	72	1.333	862367	863042	862970	862741	0.861	≤ 0.001	0.519 (0.405-0.582)
4 Class	-422567	93	1.302	845321	846192	846099	845804	0.867	≤ 0.001	0.420 (0.318-0.516)
5 Class	-418019	114	1.355	836266	837334	837220	836858	0.864	≤ 0.001	0.357 (0.252-0.432)
6 Class	-415283	135	1.461	830836	832101	831966	831537	0.862	≤ 0.001	0.312 (0.178-0.392)
7 Class	-413365	156	1.502	827043	828505	828349	827853	0.860	≤ 0.001	0.281 (0.157-0.369)
8 Class	-411776	177	1.531	823905	825564	825387	824824	0.838	≤ 0.001	Na
Factor mixture models based on 9 factors (Model 3)										
1 Class	-354910	36	1.164	709893	710230	710194	710080	Na	Na	Na
2 Class	-346798	55	1.209	693706	694222	694167	693992	0.592	≤ 0.001	Na
3 Class	-344429	74	1.309	689007	689700	689626	689391	0.679	≤ 0.001	0.570 (0.424-0.773)
4 Class	-342199	93	1.365	684584	685456	685363	685067	0.596	≤ 0.001	0.528 (0.401-0.675)
5 Class	-340702	112	1.317	681629	682678	682566	682210	0.597	≤ 0.001	0.507 (0.381-0.656)
6 Class	-339783	131	1.314	679829	681057	680926	680509	0.598	≤ 0.001	0.473 (0.360-0.647)
7 Class	-339050	150	1.406	678399	679805	679655	679178	0.595	≤ 0.001	0.434 (0.307-0.545)
8 Class	-338393	169	1.431	677123	678707	678538	678001	0.602	≤ 0.001	Na
Latent profiles models based on 9 factors controlling for the higher-order factor (Model 4)										
1 Class	-374989	36	1.150	750049	750387	750351	750236	Na	Na	Na
2 Class	-368158	55	1.160	736426	736942	736887	736712	0.500	≤ 0.001	Na
3 Class	-365862	74	1.413	731873	732566	732492	732257	0.537	≤ 0.001	0.619 (0.506-0.693)
4 Class	-364456	93	1.907	729099	729970	729877	729582	0.537	≤ 0.001	0.546 (0.471-0.592)
5 Class	-363067	112	1.343	726359	727408	727296	726940	0.546	≤ 0.001	0.499 (0.380-0.573)
6 Class	-361983	131	1.363	724228	725456	725325	724909	0.567	≤ 0.001	0.453 (0.332-0.559
7 Class	-361166	150	1.351	722632	724038	723888	723411	0.574	≤ 0.001	0.429 (0.318-0.599)
8 Class	-360375	169	1.359	721089	722673	722504	721967	0.585	≤ 0.001	Na

Note. LL = Model loglikelihood; #fp = number of free parameters; SF: scaling factor of the robust Maximum Likelihood estimator; AIC = Akaïke Information Criterion; CAIC = Consistent AIC; BIC = Bayesian Information Criterion; ABIC = sample-size Adjusted BIC; BLRT = Bootstrap Likelihood Ratio Test