Multivariate Behavioral Research, 44:764–802, 2009

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273170903333665

Psychology Press
Taylor & Francis Group

Doubly-Latent Models of School Contextual Effects: Integrating Multilevel and Structural Equation Approaches to Control Measurement and Sampling Error

Herbert W. Marsh
University of Oxford, UK

Oliver Lüdtke
Max Planck Institute for Human Development, Berlin, Germany

Alexander Robitzsch
Institute for Educational Progress, Berlin, Germany

Ulrich Trautwein
Max Planck Institute for Human Development, Berlin, Germany

Tihomir Asparouhov
Muthén & Muthén Inc., Los Angeles

Bengt Muthén
University of California, Los Angeles

Benjamin Nagengast
University of Oxford, UK

This article is a methodological-substantive synergy. Methodologically, we demonstrate latent-variable contextual models that integrate structural equation models

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

(with multiple indicators) and multilevel models. These models simultaneously control for and unconfound measurement error due to sampling of items at the individual (L1) and group (L2) levels and sampling error due the sampling of persons in the aggregation of L1 characteristics to form L2 constructs. We consider a set of models that are latent or manifest in relation to sampling items (measurement error) and sampling of persons (sampling error) and discuss when different models might be most useful. We demonstrate the flexibility of these 4 core models by extending them to include random slopes, latent (single-level or cross-level) interactions, and latent quadratic effects.

Substantively we use these models to test the big-fish-little-pond effect (BFLPE), showing that individual student levels of academic self-concept (L1-ASC) is positively associated with individual level achievement (L1-ACH) and negatively associated with school-average achievement (L2-ACH)—a finding with important policy implications for the way schools are structured. Extending tests of the BFLPE in new directions, we show that the nonlinear effects of the L1-ACH (a latent quadratic effect) and the interaction between gender and L1-ACH (an L1 $\times$ L1 latent interaction) are not significant. Although random-slope models show no significant school-to-school variation in relations between L1-ACH and L1-ASC, the negative effects of L2-ACH (the BFLPE) do vary somewhat with individual L1-ACH.

We conclude with implications for diverse applications of the set of latent contextual models, including recommendations about their implementation, effect size estimates (and confidence intervals) appropriate to multilevel models, and directions for further research in contextual effect analysis.

Complex substantive issues require sophisticated methodologies—this is the essence of substantive-methodological synergies (Marsh & Hau, 2007). Here we explore applications of the integration of structural equation modeling (SEM) and multilevel modeling (MLM) to the issue of contextual analysis, seeking to control simultaneously measurement error due to sampling of items and sampling error due to sampling of persons. We apply these latent contextual models to extend tests of effects of attending academically selective schools on academic self-concept. This research is substantively important in relation to theory (self-concept formation and frame of reference effects) and has important policy implications about tracking, ability grouping, and the organization of school systems.

OVERVIEW OF METHODOLOGICAL AND SUBSTANTIVE ISSUES: A SUBSTANTIVE-METHODOLOGICAL SYNERGY

Methodological Focus: Contextual Effects

Broadly, contextual studies evaluate whether group-level (L2) characteristics (e.g., family, peer group, classroom, school, workplace, country) contribute to out-

comes beyond what can be explained by individual-level (L1) characteristics. In many diverse disciplines, L2 constructs are based on the aggregation of variables from L1. This general strategy and related value-added models are at the heart of research in education (school effectiveness studies and associated league tables, value-added models, classroom/school climate studies), organizational psychology studies, family research, sociology, and increasingly in medical research in which related value added-models are being used to evaluate health setting effectiveness (e.g., Croon & van Veldhoven, 2007; Iverson, 1991). Contextual models are applied particularly widely in industrial, organizational, and educational psychology where researchers seek to unconfound the effects of individuals (e.g., workers, students) from those of the organizations or schools to which the individuals belong (e.g., Bliese, 2000; Bliese, Chan, & Ployhart, 2007; LaHuis & Ferguson, 2009; also see Kozlowski & Klein, 2000). In fact, the issues are central to any area of research in which individuals interact with other individuals in a group setting, leading Iverson (1991) to conclude, "This range of areas illustrates how broadly contextual analysis has been used in the study of human behavior" (p. 11). In each case, L1 attributes of persons (e.g., achievement or climate ratings in schools, employee satisfaction or productivity in business settings, family functioning, and health in hospital settings) are aggregated to form L2 constructs. A contextual effect is said to occur when there are group-level effects for subsequent outcomes beyond what can be explained in terms of the individual characteristics. Although contextual studies have a long history across many social science disciplines, recent methodological advances offer opportunities to better understand juxtapositions between individual and group effects.

Multilevel Latent Model of Contextual Effects

There is widespread application of (single-level) SEMs with multiple indicators of individual level constructs and of MLM studies in which constructs at each level are based on (single) manifest indicators of each construct. Nevertheless, progress has been slow in integrating these two dominant analytical approaches into a single framework in a way that they can be easily implemented in applied research—the focus of this investigation. Early developments (e.g., Goldstein & McDonald, 1988; McDonald, 1993, 1994; also see Goldstein, 2003) laid the foundation for important advances but they were not easily implemented with existing software (e.g., McDonald, 1994). B. O. Muthén (1989, 1994) demonstrated multilevel SEM applications using his partial maximum likelihood estimator and subsequently implemented a full information likelihood estimation procedure. Bovaird (2007) reviewed developments and statistical packages in this area but recommended Mplus as being particularly versatile for all forms of latent-variable modeling, including the integration of SEM and MLM as illustrated here (see also Skrondal & Rabe-Hesketh, 2004). Rabe-Hesketh, Skrondal,

and Pickles (2004) made a similar point, arguing, "A synthesis of both methods, namely multilevel structural equation models, is required when the units of observation form a hierarchy of nested clusters and some variables of interest cannot be measured directly but are measured by a set of items or fallible instruments" (p. 168), demonstrating how this can be incorporated into their generalized linear latent and mixed models (GLLAMM) framework.

Lüdtke et al. (2008, 2009; also see Mehta & Neale, 2005) described a multilevel latent variable model that corrected the bias in parameter estimates of contextual analysis due to the sampling error associated with aggregating L1 constructs to form L2 constructs. We refer to the Lüdtke et al. (2008) model as a partial correction model in that it corrects for sampling error in the aggregation of L1 constructs to form L2 constructs but not measurement error due to the sampling of items (manifest-variable, latent-aggregation M2 in Figure 1). Here, we extend their work by applying a doubly latent approach that additionally controls for measurement error at L1 and L2 through the use SEMs based on multiple indicators of each (latent) factor. Based on these different approaches to correction for sampling and measurement error, we consider a set of contextual models (see Figure 1; also see Lüdtke et al., 2009) that are (a) manifest or

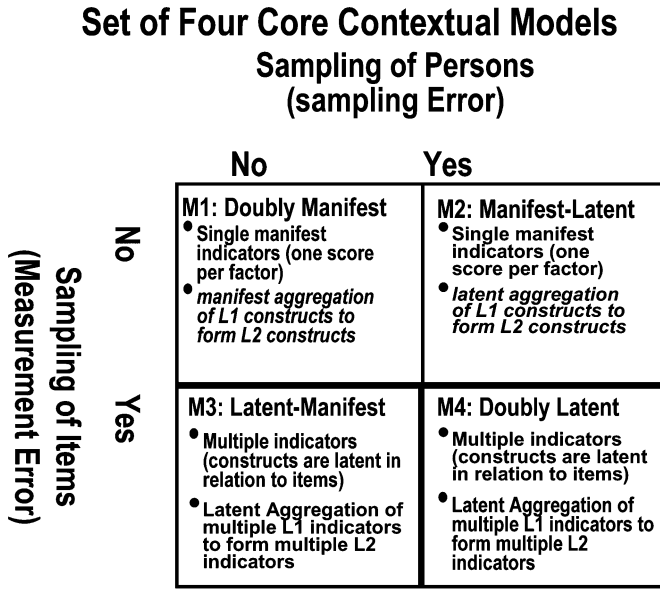

FIGURE 1 Set of four contextual models designed to control for measurement error (associated with sampling of items) and for sampling error (associated with sampling of people).

latent variable in relation to sampling items (i.e., use multiple indicators to control for measurement error, a traditional focus of confirmatory factor analysis (CFA) and SEM studies) and (b) manifest or latent in relation to the aggregation of L1 constructs to form L2 constructs (i.e., use latent aggregation to correct for sampling error in the aggregation of L1 constructs to form L2 constructs, a traditional focus of MLM studies). Traditional contextual models are often doubly manifest (manifest variable, manifest aggregation; see doubly manifest M1 in Figure 1), controlling neither measurement error nor sampling error. Although unreliability has many potential components of error (as emphasized, e.g., in generalizability theory; see Kane, Gillmore, & Crooks, 1976), here we emphasize measurement error (associated with sampling items, a traditional focus of CFA/SEM studies) and sampling error (associated with sampling of persons with L2 groups).

Significantly, the four central models (Figure 1; also see the description of models in the Methods section) can easily be implemented with existing software (see Appendix). Here our focus is on the doubly latent contextual model (latent variable, latent aggregation, M4 in Figure 1 and Appendix) with multiple indicators at L1 and L2 (controlling for unreliability due to measurement error at both levels) and a correction for sampling error in the aggregation from L1 to L2. However, it is also possible to test a latent-manifest (i.e., latent variable with manifest aggregation) approach in which multiple indicators are used to infer L1 and L2 constructs but without a control for sampling error, which might be appropriate when sampling error is zero (i.e., L2 constructs are based on all possible L1 units in each group). Partial correction models (M2 and M3 in Figure 1) are also useful in applied studies based on small *N*s where the full correction (doubly latent) model might have convergence problems or be unstable. Simulation research (Lüdtke et al., 2008, 2009) shows that under conditions of small *N* at L1 and L2, more complex models are prone to nonconvergence and can result in highly unstable estimates even when solutions converge to fully proper solutions. In this regard, one of the partial correction models (M2 or M3) might be more accurate—despite the bias—than an unbiased but unstable estimate based on M4. For this reason, it is also useful to compare the results from all four models.

Although the basic four contextual models (Figure 1) have random intercepts with fixed slopes, we extend these models to estimate random slopes (i.e., allow the slopes between L1 constructs to differ in each L2 unit) and predictors of the variation in the slopes. Further demonstrating their flexibility, we also extend these contextual models to include latent interactions (between latent L1 constructs or cross-level interactions between latent L1 and L2 constructs) or latent quadratic effects. We compare the results of the doubly latent contextual model (M4 in Figure 1) with other models in the set and demonstrate how all models can be implemented with commercially available software (see Appendix).

Substantive Focus: Big-Fish-Little-Pond Effect (BFLPE) of Selective School on Self-Concept

There is a revolution sweeping psychology (e.g., Seligman & Csikszentmihalyi, 2000) that emphasizes a positive psychology focusing on how healthy, normal, and exceptional individuals can get the most from life. Positive self-beliefs are at the heart of this revolution (Bruner, 1996; Marsh & Craven, 2006; Seligman & Csikszentmihalyi, 2000). The importance of self-concept and related constructs is highlighted by the frequency with which their enhancement is identified as a major focus in diverse settings, including education, child development, mental and physical health, social services, industry, and sport/exercise. Not only do self-concepts reflect previous life experience but they also facilitate future life successes. Persons with a positive self-concept in a particular area are likely to make critical decisions, engage in appropriate behavior, and pursue positive and challenging life experiences in that area. These critical life decisions and behaviors are likely to reinforce those positive self-concepts and lead to further accomplishments. If significant others or organizations inadvertently undermine self-concepts in an attempt to improve desired outcomes, they are likely to undermine the very outcomes they seek to maximize. Empirical support for these claims is particularly strong in educational settings where there is good evidence for the reciprocal effects of academic self-concept and achievement (e.g., Marsh, 2007; Marsh & Craven, 2006).

Our substantive focus is the widely supported big-fish-little-pond effect (BFLPE), a classic contextual effect in which the effect of individual student achievement (L1-ACH) on academic self-concept (L1-ASC) is positive, but the corresponding effect of group-average (school or classroom) achievement (L2-ACH) is negative (see Marsh, 2007; Marsh & Craven, 2002; Marsh & Hau, 2003; Marsh, Seaton, et al., 2008). Although applied primarily to educational research, the historical and theoretical underpinnings of the BFLPE theoretical model come from a variety of disciplines, including psychophysics, social psychology, organizational psychology, and sociology (see Marsh, 2007; Marsh, Seaton, et al., 2008). The BFPLE is domain specific in that it has been largely supported in relation to academic self-concept in academic settings but not for global self-esteem or other nonacademic components of self-concept (Marsh, 1987, 2007; Marsh, Chessor, Craven, & Roche, 1995; Marsh & Craven, 2006; Marsh, Seaton, et al., 2008). The BFLPE is a robust, long-lasting contextual effect that generalizes across diverse research settings, levels of education, and cultures from all over the world (Marsh, Seaton et al., 2008). From a policy perspective, the BFLPE provides an alternative, contradictory perspective to educational policy on the placement of students in special education settings, one that is being enacted in many countries throughout the world.

Historically (e.g., Marsh, 1984, 1987, 1991; Marsh & Parker, 1984), BFLPE research was based on single-level models that are unacceptable by current standards. Marsh (1984) used a single-level multiple regression based on manifest scores, using a small number of schools. Subsequent applications (Marsh, 1987, 1991) again used single-level multiple regression with manifest variables but included more schools and a crude design effect to compensate for the clustered sampling. Marsh (1994) then applied a single-level SEM in which key constructs were measured with multiple indicators. Although there was only one indicator of L1-ACH available, each school was divided into random thirds. School averages based on the random thirds were used as separate indicators of L2-ACH, providing a crude control for sampling error in the L2-ACH that foreshadowed latent contextual models considered here.

The first BFLPE study to use a true MLM (Marsh & Rowe, 1996) was a two-level MLM based on manifest indicators and manifest aggregation (e.g., doubly manifest M1 in Figure 1). Subsequent BFLPE studies (e.g., Lüdtke, Köller, Marsh, & Trautwein, 2005; Marsh, Kong, & Hau, 2000; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2006; Marsh, Trautwein, Lüdtke, & Köller, 2008) have typically been based on this manifest-manifest contextual model (M1 in Figure 2A) in which L2 was either school or class, depending on the design of the study. Marsh and Hau (2003) subsequently applied a three-level model (Level 1 = students, Level 2 = schools, Level 3 = countries) with the Organisation for Economic Co-operation and Development (OECD)/Programme for International Student Assessment (PISA) study, demonstrating the cross-national generalizability of the BFLPE. In a recent BFLPE study, with a particularly complex factor structure (19 constructs inferred from multiple indicators measured over an 8-year period), Marsh and O'Mara (2008) implemented the "complex design" option in Mplus, which only corrects standard errors for the dependencies arising from the nesting of students within classrooms instead of an MLM. Apparently, there have been no BFLPE studies implementing either partial correction models (M2 and M3 in Figure 1) or doubly latent models (M4 in Figure 1) that more clearly separate measurement error (sampling of items) and sampling error (sampling of persons).

Within the BFLPE literature (Marsh & Hau, 2003; Marsh, Seaton, et al., 2008) there is concern as to whether the BFLPE (the negative effect of L2-ACH) interacts with individual student characteristics—particularly student ability. Theoretically there is controversy with some suggesting that the BFLPE should be largest for the lowest achieving students and smallest (or even reversed) for the highest achieving students, whereas the theoretical perspective proposed by Marsh and colleagues suggests that the size of the BFLPE should be similar for all students (although tests of this might be complicated by scaling issues, including floor and ceiling effects). Substantively, this is a critical issue in that it fundamentally influences the way in which school systems should be designed. Empirical research

has been largely consistent with Marsh's theoretical prediction that the BFLPE generalizes over individual student levels (e.g., Marsh, Seaton, et al., 2008). Observed interactions tend to be small and not even consistent in their direction. However, although interaction effects are widely posited in theoretical and applied research, they are typically weak and not replicable, due, at least in part, to inherent problems of unreliability of interactions and manifest variables upon which they are based. For single-level analyses, recent developments in latent interactions (Klein & Moosbrugger, 2000; Marsh, Wen, & Hau, 2004, 2006; Marsh et al., 2007) now allow researchers to control L1 measurement error that has been such a serious problem in interaction effects based on manifest variables. However, in BFLPE research these developments have not yet been extended to multilevel cross-level interactions between individual (L1) and group (L2) characteristics that are critical in contextual studies (e.g., Marsh, 1987, 1991; Marsh, Martin, & Cheng, 2008; Marsh, Seaton, et al., 2008). More generally, in the present investigation we demonstrate how these MLMs can be addressed by applied researchers using easily implemented extensions of the set of contextual models (Figure 1) with readily available commercial software (Appendix).

METHODS

Sample

Data are part of the large, ongoing German study (Transformation of the Upper Secondary School System and Academic Careers) conducted by the Max Planck Institute for Human Development, Berlin, and the Institute for Psychology II at the University of Erlangen-Nuremberg. Up to 40 randomly selected students from 149 randomly selected upper secondary schools in one German state were representative of upper secondary schools. Students ($N = 4{,}475$, 45% males) were in their final year of upper secondary schools (typically ages 17–19). Two trained research assistants administered materials in each school in February to May 2002. The participation rate was more than 99% at the school level and 80.2% at the student level.

Materials

Math self-concept was measured with the German adaptation (Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2006; Schwanzer, Trautwein, Lüdtke, & Sydow, 2005) of the Self-Description Questionnaire SDQ III (Marsh & O'Neill, 1984). Preliminary results based on the English language version of the instrument, independent translations by four German researchers with English as a second language, assistance of a professional translator, and extensive pilot testing

resulted in a short German instrument with four items per scale with a 4-point (*disagree–agree*) response format. The selected items met two conditions: (a) they focused on competency (e.g., "I'm good at mathematics") rather than on the affective relation to the specific domain (e.g., "I like mathematics") and (b) they had the highest factor loadings in pilot studies. In support of the construct validity of this measure of math self-concept, Marsh, Trautwein, et al., (2006) demonstrated that it was substantially correlated with math achievement test scores, math school-based performance, and coursework selection in mathematics. The three indicators of mathematics achievement were based on the original items from the Third International Mathematics and Science Study (Baumert, Bos, & Lehmann, 2000).

Statistical Models and Analyses

In the following we describe the contextual analysis models as they are applied to the BFLPE (also see path diagrams in Figure 2). In this study, we have a two-level structure with students nested within schools and an individual-level construct X_{ij} (achievement of student i in school j) predicting the dependent construct Y_{ij} (self-concept of student i in school j). Both the independent and the dependent constructs are measured by multiple indicators, $k = 1, \ldots, K$ (for achievement) and $l = 1, \ldots, L$ (for self-concept).

For M1, at the student level (Level 1) the covariate is calculated by $\overline{X}_{\bullet ij} = \frac{1}{K}\sum_{k=1}^{K} X_{kij}$ averaging across the K items for each student i in school j. Correspondingly, at the school level (Level 2) the covariate is calculated by summing across the K items and the n_j persons in each group j: $\overline{X}_{\bullet\bullet j} = \frac{1}{K \cdot n_j}$ $\sum_{i=1}^{n_j}\sum_{k=1}^{K} X_{kij}$. In a similar vein the dependent variable is calculated by $\overline{Y}_{\bullet ij} = \frac{1}{L}\sum_{l=1}^{L} Y_{lij}$. The structural equation (also see M1 in Figure 2) is specified as follows:

$$\overline{Y}_{\bullet ij} = \beta_0 + \beta_1(\overline{X}_{\bullet ij} - \overline{X}_{\bullet\bullet j}) + \beta_2\overline{X}_{\bullet\bullet j} + \delta_{0j} + \varepsilon_{ij}. \quad (1)$$

In this structural equation β_0 is the grand-mean intercept, β_1 is the within-group regression coefficient describing the relationship within schools and β_2 is the between-group regression coefficient that indicates the relationship between the schools means (Cronbach, 1976), and ε_{ij} and δ_{0j} are normally distributed (with an expected value of zero) and uncorrelated residuals at L1 and L2. A contextual effect occurs if β_2 is significantly different from β_1. Because Model 1 is manifest in relation to combining items (i.e., ignores measurement error) and manifest in relation to aggregation from L1 to L2 (i.e., ignores sampling error), we label this the *doubly manifest* approach to estimating group effects in the MLM. As can be seen (Figure 2), the doubly manifest approach uses observed scores for achievement at the within and the between level (indicated by squares).

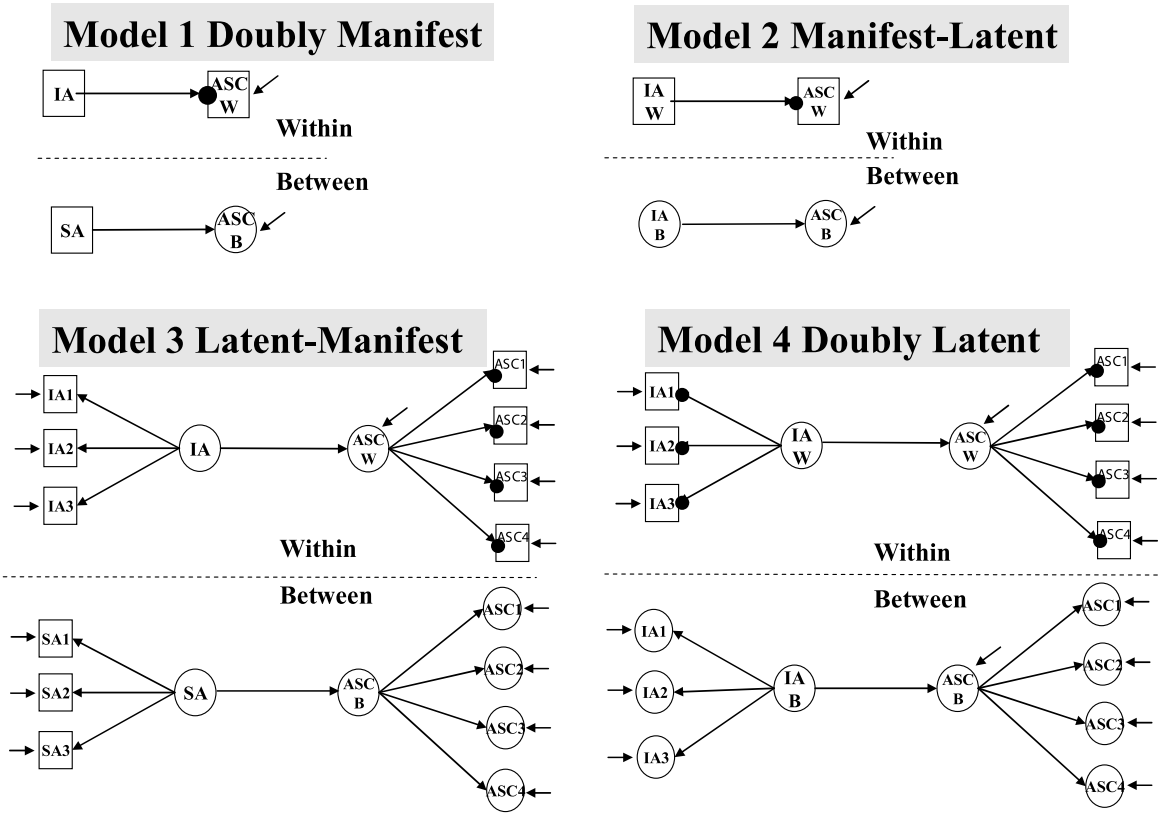

FIGURE 2A Set of four multilevel latent contextual models (see Figure 1) that are latent or manifest in relation to sampling items (and correction for measurement error) and latent or manifest in relation to sampling students (and correction for sampling error).

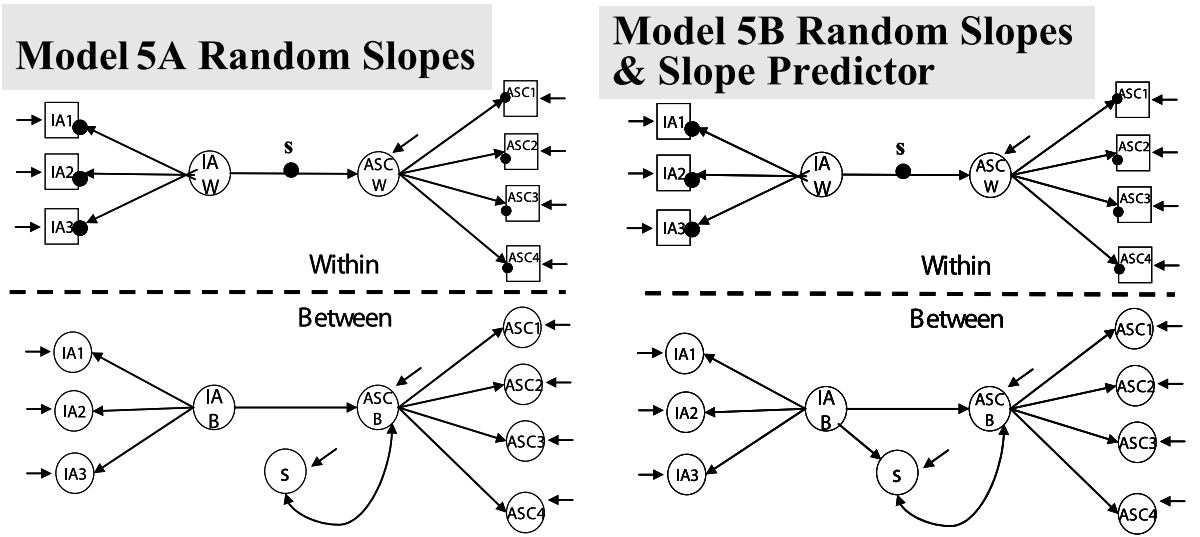

FIGURE 2B Model 5A (random slopes [s], school-to-school variation in the effect of individual achievement on individual academic self-concept) and Model 5B (slope predictor; effect of latent school-level achievement on slope, a cross-level interaction).

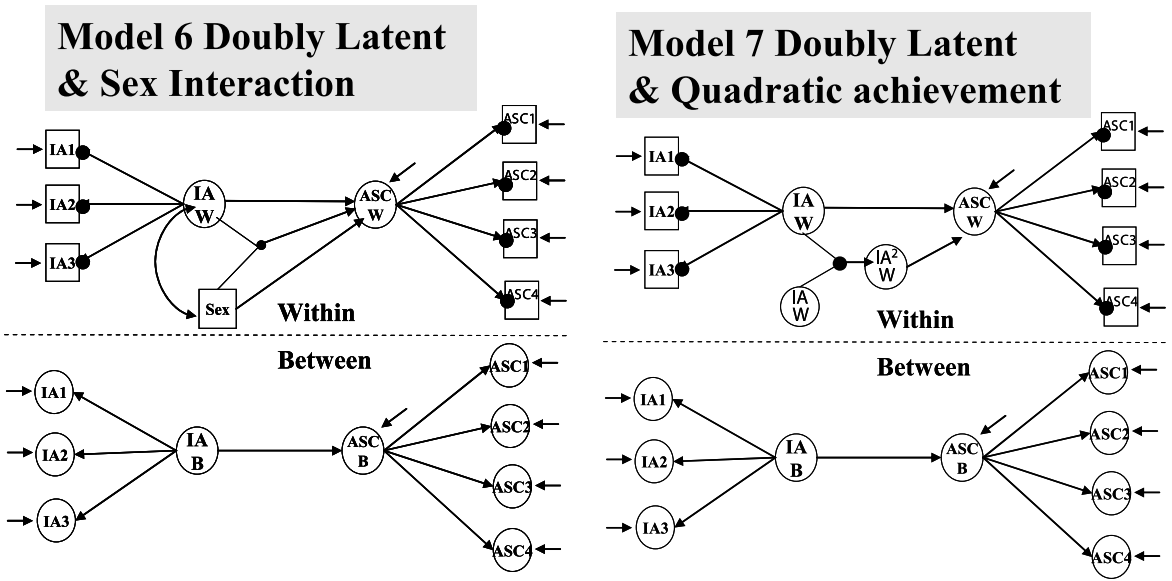

FIGURE 2C Model 6 (Sex and its interaction with latent individual achievement added to M4) and Model 6 (quadratic effect of latent individual achievement added to M4).

Figure 2 Note. IA = individual achievement (IA1–IA3 represent the multiple indicators of IA); SA = school-average achievement (SA1–SA3 represent the multiple indicators of IA); ASC = academic self-concept (ASC1–ASC4 represent the multiple indicators of ASC); W = within (student level); B = between (school level). Straight (one-directional) arrows represent paths, whereas curved (double-headed) arrows represent covariation. Circles indicate latent variables; squares indicate observed (manifest) variables. Within and between levels are separated by a dashed line. A dot at the end of the within (Level 1) regression (either from dependent on independent variables or from observed variables on latent factors) indicates that the corresponding intercept is random at the between level, representing the latent aggregation process. Thus, both single and multiple indicators at the between level (Level 2) are represented by squares if they are manifest (based on school-average values that are formed outside of the model) and circles if they are latent (formed through the latent aggregation process as part of the model estimation process). All between-level random intercepts are represented as latent variables. Interaction terms are represented by paths from the two interacting variables joining at a dot and one path leading from the dot representing the interaction effect leading to the dependent variable.

One problematic aspect of the manifest contextual analysis model is that the observed school average $\overline{X}_{\bullet\bullet j}$ might be a highly unreliable measure of the unobserved school average because only small numbers of L1 students are sampled from each L2 school (O'Brien, 1990). Lüdtke et al. (2008) introduced a multilevel latent covariate approach (see M2 in Figure 2) that takes into account sampling error when estimating group effects (see also Croon & van Veldhoven, 2007). In this approach the true group mean is considered an unobserved latent variable U_{xj} that is measured with a certain amount of precision by the group mean of the observed data (Asparouhov & Muthén, 2007). The precision is given by $\frac{\tau_x^2}{\tau_x^2+(\sigma_x^2/n_j)}$ where τ_x^2 is the variance between groups and σ_x^2 is the variance within groups. In the literature on reliability of multilevel data (Bliese, 2000) this measure is also sometimes called the ICC(2) and is used to determine the reliability of aggregated individual level data (e.g., the observed school average $\overline{X}_{\bullet\bullet j}$) in terms of sampling only a finite number of L1 units from each L2 unit. Thus, it can be interpreted as the reliability of the group mean in relation to sampling error. In most cases, the mean group size can be entered for n_j if not all groups are of the same size (see Searle, Casella, & McCulloch, 1992, on how to deal with pronounced differences in group size).

As is typical within structural equation modeling (SEM), the estimate of the group-level coefficient is then corrected for the unreliable assessment of the latent group mean by the observed group mean. The structural equation for the Model 2 (also see Model 2 in Figure 2) is given as follows:

$$\overline{Y}_{\bullet ij} = \beta_0 + \beta_1(U_{\bullet ij} - U_{\bullet\bullet j}) + \beta_2 U_{xj} + \delta_{0j} + \varepsilon_{ij}, \tag{2}$$

where $U_{\bullet ij}$ and $U_{\bullet\bullet j}$ are latent variables corresponding to the manifest variables $\overline{X}_{\bullet ij}$ and $\overline{X}_{\bullet\bullet j}$ in Equation 1, and U_{xj} is a latent variable that is corrected for sampling error. However, in the M2, measurement error due to the sampling of items at L1 and L2 distorts the estimation of the corresponding regression coefficients at β_1 and β_2. This also becomes apparent in the path diagram in Figure 2 in which student achievement is a latent variable at L2 but a manifest at L1. Because the M2 does not take into account measurement error we refer to it as the *manifest-latent* approach, manifest in that it starts with scale scores or single indicators and latent in that it controls for sampling error.

In M3 both the independent and the dependent variables are measured by multiple indicators. By extending the classical CFA model, a multilevel CFA with a within-group and between-group measurement model can be defined (see B. O. Muthén, 1991). Using this approach, single indicators of the dependent variable Y can be decomposed as follows (also see Model 3 in Figure 2):

$$Y_{lij} = \mu_{ly} + \lambda_{ly,W} U_{yij} + R_{lyij} + \lambda_{ly,B} U_{yj} + R_{lyj}; \quad l = 1, \ldots, L, \tag{3}$$

where $\lambda_{ly,W}$ are the within-factor loadings, $\lambda_{ly,B}$ are the between-factor loadings, R_{lyij} are the residuals at Level 1, and R_{lyj} are the residuals at Level 2. U_{yij} and U_{yj} are the unobserved true scores at L1 and L2. In classical test theory this represents a congeneric measurement model because the factor loadings are allowed to vary across indicators. Although the independent variable is also measured by multiple indicators ($k = 1, \ldots, K$), in the latent-manifest covariate approach the latent factor at L2 is based on manifest indicators that are the result of manifest aggregations of the observed indicators to the group level: $\overline{X}_{k\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{kij}$. Thus, the measurement model at Level 1 is

$$X_{kij} - \overline{X}_{k\bullet j} = \mu_{kx} + \lambda_{k,W} U_{xij} + R_{xkij}; \quad k = 1, \ldots, K.$$

The measurement model at Level 2 is

$$\overline{X}_{k\bullet j} = \mu_{kx} + \lambda_{k,B} U_{xj} + R_{xkj}; \quad k = 1, \ldots, K.$$

The resulting structural model is

$$U_{yij} = \beta_0 + \beta_1 U_{xij} + \beta_2 U_{xj} + \delta_{0j} + \varepsilon_{ij}.$$

For example, using this approach the L1 achievement factor would be based on responses to the three L1 achievement indicators and the L2 (school-average) achievement factor is based on the responses to simple (manifest) school-average values of the corresponding L1 indicators (see path diagram in Figure 2 where L2 indicators of achievement are represented as squares, representing manifest-aggregate variables, rather than circles, representing latent-aggregate variables). Hence, L2 achievement is latent in the sense that it is based on multiple indicators. However, it is manifest in relation to aggregation from L1 to L2 in the sense that it does not correct for sampling error associated due to sampling of individuals (i.e., with within-class variation in L1 achievement scores as does Model 2). Hence, we refer to M3 as the *latent-manifest* approach. This reflects the fact that it is latent in terms of the measurement model at both levels but that it is manifest in terms of not taking into account the sampling error.

M4 (Figure 2) is doubly latent in that it takes into account measurement error at L1 and L2 (based on multiple L1 indicators) and L2 sampling error due to the aggregation from L1 to L2. In this sense it builds on the *latent-manifest* (M3) approach (that has latent L1 and L2 factors in which measurement error was controlled based on consistency among multiple indicators but did not incorporate corrections of L2 sampling error based on consistency among students within each class) and the *manifest-latent* (M2) approach (that did not control for sampling error in the aggregation from L1 to L2). Hence, in addition to the decomposition of the outcome variable given in Equation 3, M4 assumes

the following equation for the indicators of the predictor variables:

$$X_{kij} = \mu_{xk} + \lambda_{kx,W}U_{xij} + R_{xkij} + \lambda_{kx,W}U_{xj} + R_{xkj}; \quad k = 1, \ldots, K. \tag{4}$$

The resulting structural model is again

$$U_{yij} = \beta_0 + \beta_1 U_{xij} + \beta_2 U_{xj} + \delta_{0j} + \varepsilon_{ij}.$$

As can be seen in Figure 2, in the doubly latent M4 the indicators of achievement at L2 as well as the factor at L1 are considered latent variables (i.e., are circles rather than squares). This type of model was also previously described by Rabe-Hesketh, Skrondal, and Pickles (2004) as a special case of their GLLAMM framework (see Equation 19, p. 181; see also McDonald, 1993, 1994).

In M5A we extend M4 by allowing the within-group regression coefficients β_1 to vary across the schools, a random slope or random coefficient model (see M5A in Figure 2):

$$U_{yij} = \beta_0 + \beta_1 U_{xij} + \beta_2 U_{xj} + \delta_{0j} + \delta_{1j} U_{xij} + \varepsilon_{ij}, \tag{5A}$$

where δ_{1j} represents the normally distributed (with an expected value of zero) deviations from the average slope β_1, and δ_{0j} and δ_{1j} are allowed to covary among schools. In the next step, the interaction $U_{xij} \cdot U_{xj}$ between individual achievement and school-average achievement is added to explain variation in slopes across schools (M5B in Figure 2):

$$U_{yij} = \beta_0 + \beta_1 U_{xij} + \beta_2 U_{xj} + \beta_3 U_{xij} \cdot U_{xj} + \delta_{0j} + \delta_{1j} U_{xij} + \varepsilon_{ij}. \tag{5B}$$

Tests for the cross-level interaction in M5B are appropriate even when the variance component in M5A is nonsignificant (e.g., LaHuis & Ferguson, 2009), but the size of the variance component is an important consideration in the interpretation of the cross-level interaction.

In M6 we added to M4 the sex of the student and the interaction between individual achievement and sex (also see M6 in Figure 2):

$$U_{yij} = \beta_0 + \beta_1 U_{xij} + \beta_2 U_{xj} + \beta_3 SEX_{ij} + \beta_4 U_{xij} \cdot SEX_{ij} + \delta_{0j} + \varepsilon_{ij}. \tag{6}$$

Finally, in M7 (Figure 2) quadratic individual achievement is added to M4 to evaluate possible nonlinear relations between self-concept and achievement:

$$U_{yij} = \beta_0 + \beta_1 U_{xij} + \beta_2 U_{xj} + \beta_3 U_{xij}^2 + \delta_{0j} + \varepsilon_{ij}. \tag{7}$$

All statistical analyses were conducted with Mplus (Version 5.2). For estimating parameters in MLMs, Mplus uses a general approach that is based on an

accelerated expectation maximization (EM) algorithm that provides maximum likelihood estimates for two-level structural equation models with missing data (Asparouhov & Muthén, 2003; also see Lüdtke et al., 2008). This general model incorporates random coefficients and integrates the modeling frameworks of MLMs and SEMs. It also provides robust estimates of the asymptotic covariance of the maximum likelihood estimates and the chi-square test. These models can be fitted with the approach described by Lee and Poon (1998) that handles only random intercept models, but Mplus takes a more general approach with random slopes. The Mplus approach does not require an assumption of normality with the maximum likelihood robust (MLR) estimator because it implements nonnormality robust *SE* calculations. More important, it is now relatively easy to estimate random slopes models in an SEM framework (see Equations 5A and 5B), a problem that has plagued SEM researchers (Kaplan, 2000).

In MLMs it is typical to distinguish between group mean centering and grand mean centering (Enders & Tofighi, 2007; Kreft, de Leeuw, & Aiken, 1995). In all models considered here, we used group mean centering; the mean of school j is subtracted from the score of student i in school j; $(\overline{X}_{\bullet ij} - \overline{X}_{\bullet\bullet j})$. As discussed earlier, a contextual effect in the group mean centering model is present if the L2 between-school regression coefficient is significantly different from the L1 within-school regression coefficient. A test of this hypothesis was accomplished by calculating an additional parameter—the difference between the regression weights assigned to individual and L2-ACH—that is a direct estimate of the BFLPE (see Appendix). Although this could easily be done by hand, the important advantage of the approach used here is that it also provides a standard error of the estimate. Within this framework, it is straightforward to incorporate interaction terms (see Equations 6 and 7) based on the latent moderated structural equation algorithm proposed by Klein and Moosbrugger (2000; also see Marsh, Wen, & Han, 2004). This procedure is based on the analysis of the multivariate distribution of the joint indicator vector and explicitly takes the specific type of nonnormality implied by latent interactions into account. The joint distribution of indicator variables is represented as a finite mixture of normal distributions.

RESULTS

We begin with a single-level model with manifest variables and manifest aggregation. We label this M0 to emphasize that it is not an MLM. However, it was an important historical basis of BFLPE studies (e.g., Marsh, 1984, 1987) and contextual studies more generally. Nevertheless, such single-level doubly manifest models are clearly inappropriate in that at least the standard errors are likely to be substantially biased and that it precludes many interesting

MLM questions. In predicting L1-ASC in M0 (see Table 1), the effect of L1-ACH (.605) is substantial. However, because of the within-group centering, the estimated effect of L2-ACH (.292) is not a direct estimate of the BFLPE of L2-ACH. Hence, we calculated an additional parameter—the difference between weights for individual and L2-ACH—that is a direct estimate of the BFLPE and

TABLE 1
BFLPE Models: Single and Multilevel Models Based on Single (Manifest) Indicators With Either Manifest or Latent Aggregation (see Figures 1 & 2)

	Model 0: Single-Level Doubly Manifest			
	Estimates	*SE*		
ASC on				
Achieve-Ind (IA)	0.605	0.014		
Achieve-School (SA)	0.292	0.025		
Resid Var ASC	0.702	0.014		
R-SQUARE ASC	0.297	0.011		
New/Additional parameters				
BFLPE	−0.313	0.028		
	Model 1: Multilevel Doubly Manifest		*Model 2: Multilevel Manifest Indicators, Latent Aggregation*	
	Estimates	*SE*	*Estimates*	*SE*
Within (W) level				
ASC-W ON ACH-W	0.605	0.014	0.605	0.014
Resid Var ASC-W	0.685	0.013	0.685	0.013
Between (B) level				
ASC-B ON ACH-B	0.301	0.029	0.262	0.033
Resid Var ASC-B	0.017	0.005	0.014	0.005
R-SQUARE				
Within ASC	0.288	0.017	0.295	0.017
Between ASC	0.572	0.096	0.515	0.116
New/Additional parameters				
BFLPE	−0.304	0.034	−0.343	0.038

Note. IA = individual achievement; SA = school-average achievement; ASC = academic self-concept; ACH = achievement; BFLPE = big-fish-little-pond effect. For the multilevel models, ASC and ACH have separate components for the within (individual student) and between (school) levels. Goodness of fit: Both M0 and M1 are saturated so that $df = 0$ and the fit is perfect. M0: Free parameters = 4; Akaike information criterion (AIC) = 30,489; Bayesian information criterion (BIC) = 30,515; Sample-size adjuted BIC (SSA-BIC) = 30,502; M1: Free parameters = 5; AIC = 11,686; BIC = 11,719; SSA-BIC = 11,703. M2: Free parameters = 5; AIC = 24,235; BIC = 24,268; SSA-BIC = 24,252. See Appendix for the MPLUS syntax for each model.

a standard error to test for its statistical significance (see syntax in Appendix). In M0, the BFLPE is −.313 and highly significant. An increase of 1 *SD* in L2-ACH (in the metric of L1-ACH) results in a decline in L1-ASC of almost one third of an *SD*.

MLM Contextual Models (Figures 1 & 2)

Doubly manifest model (M1): Single indicators and manifest aggregation. M1 (see Figures 1 & 2) is an MLM contextual model based on manifest (single-indicator) measures of L1-ACH and L2-ACH. Aggregation is manifest in that L2-ACH is a simple (manifest) average of the L1-ACHs in each school. Although this doubly-manifest model is clearly more appropriate than M0, it implicitly assumes that there is no measurement error for any of the L1 or L2 constructs and no sampling error in aggregating L1-ACH to form L2-ACH. The BFLPE (−.304) is similar in size to estimates from the single-level M0, but the standard errors—particularly for L2 constructs—are substantially larger (reflecting the multilevel structure of the data ignored in M0).

Manifest-latent contextual model (M2): Control for sampling error. M2 (see Figures 1 & 2) is the latent-manifest MLM described by Lüdtke et al. (2008) that controls sampling error in the aggregation of L1 constructs to form L2 constructs. However, because L1 and L2 constructs are manifest (based on single indicators), M2 does not control for L1 measurement error. In M2, the effect of L1-ACH on L1-ASC is .605. However, the important feature of M2 is that the L2-ACH is latent in the sense that it corrects sampling error based the latent aggregation of L1-ACH to form L2-ACH. Hence, the BFLPE estimate for M2 (−.343) is larger than for M1 (−.304).

Latent-manifest model (M3): Control for measurement error. M3 (see Table 2) is a latent-manifest MLM (see Figures 1 & 2) in which both L1 and L2 constructs are based on multiple indicators to control for measurement error. However, M3 assumes no sampling error in the aggregation of L1-ACH to form L2-ACH. For purposes of this analysis, we specified that L1 factor loadings were the same as L2 factor loadings (see Appendix). Although not necessary, this cross-level invariance facilitates interpretations of the results (see subsequent discussion of this issue). The BFLPE for M3 (−.317) is substantial and somewhat more negative than estimates-based M0 and M1 (Table 1).

Doubly manifest contextual models (M4): Control for sampling and measurement error. M4 is a doubly latent MLM (see Figure 1) that incorporates both multiple indicators of L1 constructs to control measurement error at L1 and L2 and a latent aggregation that controls sampling error in the

TABLE 2
BFLPE Models: Single and Multilevel Models Based on Single (Manifest) Indicators With Either Manifest or Latent Aggregation (see Figures 1 & 2)

	Model 3 (Latent-Manifest): Multiple Latent Indicators, Latent Aggregation		*Model 4 (Doubly Latent): Multiple Latent Indicators, Manifest Aggregation*	
	Estimates	*SE*	*Estimates*	*SE*
Within level				
ASC factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.724	0.016	0.724	0.016
Indicator 3	0.979	0.009	0.979	0.009
Indicator 4	0.845	0.011	0.845	0.011
ACH factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.966	0.009	0.966	0.010
Indicator 3	0.951	0.015	0.952	0.015
ASC-W on ACH-B	0.624	0.015	0.624	0.015
Var Ach-W	0.711	0.046	0.734	0.048
Resid Var ASC-W	0.551	0.013	0.550	0.013
Between (B) level				
ASC-B factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.724	0.016	0.724	0.016
Indicator 3	0.979	0.009	0.979	0.009
Indicator 4	0.845	0.011	0.845	0.011
ACH-B factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.966	0.009	0.966	0.010
Indicator 3	0.951	0.015	0.952	0.015
ASC-b ON ACH-b	0.307	0.028	0.266	0.034
Var ACH-b	0.238	0.028	0.208	0.027
Resid Var ASC-b	0.016	0.004	0.013	0.004
R-SQUARE				
Latent ASC-W	0.334	0.019	0.342	0.019
Latent ASC-B	0.589	0.094	0.528	0.115
New/Additional parameters				
BFLPE	−0.317	0.033	−0.358	0.040

Note. ACH = achievement; ASC = academic self-concept; BFLPE = big-fish-little-pond effect. There are four indicators for ASC factors (within and between) and three indicators for ACH factors (within and between). See Appendix for the MPLUS syntax for each model. Goodness of fit values are M3: chi-square (31) = 443; CFI = .989; TLI = .984; RMSEA = 0.053; SRMR-B = 0.030; SRMR-W = 0.016; AIC = 60,617; BIC = 60,843; SSA-BIC = 60,732; M4: chi-square (31) = 463; CFI = .988, TLI = .984; RMSEA = 0.054; SRMR-B = 0.049; SRMR-W = 0.023; AIC = 61,898; BIC = 62,105; SSA-BIC = 62,003.

aggregation process in going from L1 to L2. In this sense, M4 is fully latent and builds on partial correction models that only correct for either sampling error (M2) or measurement error (M3). As M4 is doubly latent, the BFLPE estimate (−.367) is more negative than M1 (−.304), M2 (−.317), or M3 (−.343).

Extensions of the Doubly Latent Contextual Model: Latent Interactions and Nonlinear Terms

The contextual models (Figures 1 & 2) can easily be extended to include additional effects that are relevant to substantive issues in BFLPE research. Although the extensions described here could be applied to any of the four core contextual models, we focus on extensions to M4:

- A random effects model (the extent to which the relation between L1-ACH and L1-ASC varies from school to school),
- A latent cross-level interaction (between L1-ACH and L2-ACH),
- A latent interaction between two L1 constructs (sex and L1-ACH), and
- A latent quadratic model (quadratic component of L1-ACH).

Random slopes and latent cross-level interaction between L1-ACH and L2-ACH (M5A & M5B, Table 3). Although numerous studies have evaluated whether the BFLPE varies with L1-ACH and found weak, inconsistent, or nonsignificant effects consistent with theoretical predictions, none tested this interaction with latent contextual models like those considered here. We begin by extending M4 to include random slopes—whether the effect of L1-ACH on L1-ASC varies from school to school (labeled "slope variance" in M5A and M5B). In M5A, the variance component for this slope parameter (.001, $SE = .001$) is very small, indicating that there is no significant school-to-school variation in the size of the slope. Although not reported, the random slope parameter was nonsignificant for each of the set of four contextual models (Figure 1). In M5A, the relation between the slope and intercept is also very small (−.008, $SE = .003$).

Even though the variance component for the random slope parameter is not statistically significant, we tested the extent to which this variation is predicted by L2-ACH (ACH-b in Table 3). This cross-level interaction-how the effect of L2-ACH on L1-ASC varies with L1-ACH-is small but positive and statistically significant (.118, $SE = .037$). These results indicate that the effect of L1-ACH on L1-ASC is substantial and nearly the same across schools. Nevertheless, a significant portion of this small (nonsignificant) variance can be explained by L2-ACH. Whereas the BFLPE is substantial and negative (−.387), this negative effect is offset to some extent for the very brightest students. However, even for high-achieving students, there is a substantial negative effect of L2-ACH.

TABLE 3
Doubly Latent Model 4 With Random Slopes: Without Latent Interaction (Model 5A) or With Latent Interaction (Model 5B)

	Model 5A		*Model 5B*	
	Estimates	*SE*	*Estimates*	*SE*
Within (W) level				
ASC-W factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.724	0.016	0.724	0.016
Indicator 3	0.980	0.009	0.980	0.009
Indicator 4	0.845	0.011	0.845	0.011
ACH-W factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.975	0.011	0.975	0.011
Indicator 3	0.986	0.014	0.986	0.014
Var ACH-W	0.712	0.049	0.711	0.049
Resid Var ASC-W	0.546	0.013	0.544	0.013
Between level				
ASC-B factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.724	0.016	0.724	0.016
Indicator 3	0.980	0.009	0.980	0.009
Indicator 4	0.845	0.011	0.845	0.011
ACH-B factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.975	0.011	0.975	0.011
Indicator 3	0.986	0.014	0.986	0.014
ASC-b ON ACH-b	0.292	0.035	0.261	0.034
Var ACH-b	0.214	0.027	0.213	0.027
Resid Var ASC-b	0.016	0.004	0.015	0.005
Slope mean	0.632	0.015	0.639	0.015
Slope variance	0.001	0.001	0.003	0.002
Slope with ASC-b	−0.008	0.003	−0.006	0.003
Slope ON ACH-b			0.118	0.037
New/Additional parameters				
BFLPE	−0.340	0.040	−0.378	0.039

Note. ACH = achievement; ASC = academic self-concept; BFLPE = big-fish-little-pond effect. Slope = slope of the latent (within) ASC factor on the latent (within) ACH factor. Slope has a mean (average slope over schools) and variance (or residual variance in Model 5B), although Model 5B also tests the effect of (between) latent achievement on the slope. See Appendix for the MPLUS syntax for each model. Goodness of fit values are M5A: Free Parms = 27; AIC = 62,188; BIC = 62,362; SSA-BIC = 62,276; HO value = −31,067. M5B: Free Parms = 28, AIC = 62,176; BIC = 62,357; SSA-BIC = 62,268; HO value = −31060.

Consistent with theory and previous research, the size of the BFLPE does not vary much with the ability levels of individual students.

M6: Latent interaction between sex and L1-ACH (Table 4). There are well-established, gender-stereotypic differences in L1-ASC factors that tend to be even stronger than those observed in corresponding ACH factors. Nevertheless, previous research (Marsh & Yeung, 1998) suggests that the effect of L1-ACH on L1-ASC is similar for boys and girls (i.e., does not interact with gender). In M6 (Figure 2) we add L1 terms representing the main effect of gender and its interaction with L1-ACH to M4 (see Appendix). However, neither of these added effects was statistically significant.

M7: Latent quadratic effect of L1-ACH (Table 4). A number of BFLPE studies (e.g., Marsh, Hau, & Craven, 2004; Marsh & Rowe, 1996) reported a nonlinear, quadratic component in the relation between L1-ACH and L1-ASC. In M7 (Figure 2), we added a latent quadratic term to M4 (see Appendix). However, unlike most previous research this was a truly latent quadratic term based latent L1-ACH. However, the quadratic effect of L1-ACH on L1-ACH was not statistically significant.

DISCUSSION

Substantive-Methodological Synergy

Substantively, we provided a range of different approaches to evaluate the size of the BFLPE, an important theoretical issue in self-concept research with significant policy implications. Methodologically, we described applications of a set of latent contextual models—and their extension—that has broad applicability. Taken together, these illustrate the strength of substantive-methodological synergies and the flexibility of the latent approach to contextual models. The four core contextual models (Figure 1) offer potential trade-offs in relation to bias and robustness. In particular, under many circumstances in applied research, M4 is likely to be unbiased but might be unstable whereas the other three models (M2 and M3, or even M1) are likely be biased but might be more stable (Lüdtke et al., 2008, 2009). Thus, we recommend that applied researchers evaluate all four models. When there is a clear pattern in results based on the four models, the applied researcher can have more confidence in the interpretation of the results than when any one of the models is applied. In the present investigation, for example, the BFLPE estimates were reasonably similar across the four models, but the BFLPE estimate was somewhat larger for the full-correction M4 than for

TABLE 4
Doubly Latent Model 4 With the Addition of an L1-Latent Interaction (Model 6: Gender × L1-ACH) and an L1-Latent Quadratic Effect (Model 7: L1-ACH × L1-ACH)

	Model 6		*Model 7*	
	Estimates	*SE*	*Estimates*	*SE*
Within (W) level				
ASC-W factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.724	0.016	0.724	0.016
Indicator 3	0.980	0.008	0.980	0.008
Indicator 4	0.845	0.011	0.845	0.011
ACH-W factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.975	0.011	0.975	0.011
Indicator 3	0.986	0.014	0.986	0.014
ASC-W on ACH-W	0.648	0.040	0.645	0.016
ASC-W on SEX			0.009	0.027
ASC-W on SexXACH-W	−0.007	0.023		
ASC-W on QuadACH-W			0.009	0.005
Var ACH-W	0.712	0.049	0.712	0.049
Resid Var ASC-W	0.548	0.013	0.547	0.013
Between (B) level				
ASC-B factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.724	0.016	0.724	0.016
Indicator 3	0.980	0.008	0.980	0.008
Indicator 4	0.845	0.011	0.845	0.011
ACH-B factor loadings				
Indicator 1	1.000	0.000	1.000	0.000
Indicator 2	0.975	0.011	0.975	0.011
Indicator 3	0.986	0.014	0.980	0.014
ASC-b ON ACH-b	0.258	0.035	0.257	0.035
Var ACH-b	0.213	0.026	0.212	0.027
Resid Var ASC-b	0.015	0.004	0.015	0.004
New/Additional parameters				
BFLPE	−0.390	0.055	−0.389	0.041

Note. ACH = achievement; ASC = academic self-concept; BFLPE = big-fish-little-pond effect; SexXACH-W = interaction between L1-sex and L1-ACH; QuadACH-W = quadratic component of L1-ACH. See Appendix for the MPLUS syntax for each model. M6: No Free Parms = 27; AIC = 62195; BIC = 62370; SSA-BIC = 62,284; M7: Free Parms = 26; AIC = 62,190; BIC = 62,358; SSA-BIC = 62,276.

the partial-correction M3 and M2 and smallest for M1 that corrected for neither measurement error not sampling error.

Limitations and Directions for Further Research

Applicability and robustness of estimation procedures. This investigation is based on a robust maximum likelihood estimator (L. K. Muthén & Muthén, 2006–2008). However, for complex models like those considered here the optimization algorithms may not converge to a fully proper truly optimal solution unless L1 and L2 sample sizes are large, especially for the doubly latent M4 and extensions built upon it (M5–M7). Although this was apparently not a problem in this investigation with 149 schools and 4,475 students, the doubly latent M4 in particular may not work so effectively when sample sizes are more modest—as is common in many applied contextual studies. However, increasingly sophisticated statistical packages such as Mplus have good diagnostics that warn applied researchers when convergence to a proper solution is problematic and present a number of options to address this problem-increasing the number of iterations, increasing the number of random start values, and using better start values. By substantially increasing the number of random start values, the applied researcher can evaluate how frequently the robust ML estimation procedure results in the same log-likelihoods.

Robustness of solutions—based on the sampling variability in parameter estimates—is likely to be a problem for latent contextual models. In extensive simulation studies based on the manifest-latent M3, Lüdtke et al. (2008) showed that under some conditions (small L1 and L2 sample sizes and small ICCs) the variability of the contextual effects was large and that a large number of L2 units was necessary for the asymptotic properties to hold in finite samples. The sample-size demands for the doubly latent contextual model are even greater based on further simulation studies that investigate the finite sample performance and sample-size requirements both at L1 and L2 for these models (see Lüdtke et al., 2009).

A viable option to the problem of unstable parameter estimates based on the doubly latent M4 is to consider partial correction models (M2 and M3), or even the doubly manifest M1, that are substantially less complex. Particularly when sample sizes are small, the apparently small amounts of bias introduced by these partial correction models might be less worrisome than the potentially unstable estimates (and wide confidence intervals) resulting from the more complex doubly latent M4 (see Lüdtke et al., 2009). Even with the present investigation where the sample sizes are substantial, the standard error for the BFLPE associated with the doubly latent model (.040) was larger than those associated with other less complex models (.028 to .038). We expect that this difference could be larger, depending upon the complexity of the model and

sample sizes (also see Lüdtke et al., 2009). Comparative simulations could also shed further light on the relative merits of the different MLM contextual models and address the question of which estimate comes closest to the "true" value of population generating model and the extent to which potential biases of the partial correction models (M3 and M2, or even M1) are offset by their higher efficiency in small samples (Lüdtke et al., 2008, 2009). Because of this trade-off between bias and robustness, we recommend that all four contextual models (Figure 1) should be evaluated (see earlier discussion); the most appropriate model will vary with particular circumstances of the study (e.g., ICC; the sample sizes at L1 and L2; and, perhaps, nature of the variables).

The "complex design" option offers an entirely different alternative to the problem of unstable solutions (Asparouhov & B. Muthén, 2006; B. Muthén & Satorra, 1995; also see Marsh & O'Mara, 2008). Two different traditions exist for handling issues associated with clustered samples and multistage sampling designs. The MLM approach is to make the multiple levels of the data explicit, whereas the complex design approach is to appropriately weight the parameter estimates so that the standard errors used to test the effects are appropriate. The position here is not that one approach is inherently superior to the other and indeed there are circumstances in which the results based on the two approaches should converge (B. Muthén & Satorra, 1995). Although many of the advantages of MLMs are not available when the complex design option is used, the complex design option might be particularly attractive when the sample sizes at L1 and L2 are modest. In this respect, the complex design option applied to single-level models (e.g., M0 in Table 1) provides a completely different strategy when L1 and L2 sample sizes might not be sufficient to justify the use of MLMs in contextual studies. This is an area that warrants further attention.

Finally, although beyond the scope of this investigation, Bayesian procedures with programs such as WinBUGS offer a natural framework for multilevel models that is particularly attractive when sample sizes are not substantial or the model to be estimated is very complicated (Gelman & Hill, 2007; Lee, 2007; Ntzoufras, 2009; Swaminathan & Rogers, 2008). Bayesian procedures are naturally suited to multilevel models with modest *N*s in that the sampling-based Bayesian procedures depend much less on the asymptotic theory than likelihood estimation procedures (Lee, 2007). A Bayesian framework using the WinBUGS software, which is based on Markov Chain Monte Carlo methods, can be used to integrate multilevel modeling and structural equation modeling in a very flexible way (Lee, 2007; Segawa, Emery, & Curry, 2008). However, due to the large *N*s in the present investigation, the differences between ML and Bayesian estimates would likely be very small and our focus was on the substantive interpretations of the contextual models rather than estimators per se. Nevertheless, a Bayesian option does provide an attractive alternative when samples sizes are small and more research is needed to evaluate this alternative estimator.

In summary, particularly the doubly latent model demonstrated here may not be sufficiently robust for many applied studies based on small *N*s. Although beyond the scope of this study, there is clear need for simulation studies to evaluate the robustness of procedures presented here under a variety of conditions and to explore alternative options or estimation procedures for conditions under which solutions are not stable.

Nature of the variables involved: Formative versus reflective L2 aggregation processes. Lüdtke et al. (2008) and others (e.g., Skrondal & Laake, 2001) distinguished between what they referred to as *formative* and *reflective* aggregations of L1 constructs. Although their distinction is based on a factor-analytic rationale, a related distinction is made in the organizational psychology (e.g., Bliese, 2000; Bliese et al., 2007; also see Kozlowski & Klein, 2000) between compilation (or configural) models and composition models. The main distinction between the two is that the group is the referent in the reflective aggregation process (i.e., each member of the group directly rates the L2 construct) whereas the individual is typically the referent in the formative aggregation process (i.e., the L2 aggregation is based on a group average of individual characteristics). For example, if all members of each group rated the competitive climate of the group, the aggregate would be a reflective measure, whereas if each member rated his or her own individual competitive orientation the resulting aggregate would be a formative measure. In this sense, school-average ability might more reasonably be considered a formative aggregation.

The theoretical rationale for reflective aggregations of L1 constructs is based on classical measurement theory and the domain sampling model. Group characteristics are latent, unobserved constructs that can be inferred on the basis of multiple indicators. A critical assumption is that scores for each individual within the same group reflect the same L2 construct. In this respect the group members are interchangeable (in relation to scores reflecting the L2 reflective construct) and there is a degree of isomorphism between the L1 and L2 measures. The doubly latent (M4) and manifest-latent (M2) models assume a reflective aggregation process to the construction of L2 aggregate constructs.

Formative (compilation or configural) aggregations of L1 constructs are more problematic from the perspective of L2 sampling error. The L1 measures used to construct formative aggregations are not interchangeable in the sense that individuals within the same group have different L1 true scores so that within-group heterogeneity cannot be assumed to reflect only sampling error. Lüdtke et al. (2008) give the example in which the gender of all students in each of a large number of different classes is known and each class is characterized by the percentage of females. Clearly individual students are not interchangeable in relation to gender; it is reasonable to assume that there is little or no measurement error in determination of the gender at L1 and L2, and there is

no sampling error as the sampling ratio is 100%. In this case, the manifest aggregation assumption (M1 and M3) that the L2 construct is free of sampling error may be reasonable and the assumption that within-group variance represents sampling error might be incorrect (leading to biased estimates of contextual effects based on inappropriate correction for sampling error for formative constructs). However, based on their simulation study with formative constructs, Lüdtke et al. (2008) showed that models based on the assumption of the latent aggregation approach were appropriate for formative variables when the sampling ratio was low—as in the present investigation—so that there was considerable sampling error. In this case, the latent aggregation models (M2 and M4) provide reasonable parameter estimates related to the L2 aggregated (formative) construct. However, when the sampling ratio approaches 100%—particularly when the number of individuals within each group is small—latent aggregation approaches might overestimate sampling error (based on within-group variance) and result in inflated estimates of the contextual effects. In this case, the juxtaposition of estimates based on different models would be particularly informative. Although the doubly latent M4 (and manifest-latent M2) are appropriate for purely reflective measures, more simulation research along the lines of the Lüdtke et al. (2008; also see Lüdtke et al., 2009) is needed to evaluate their appropriateness for formative measures—particularly when sampling ratios are substantial.

Equality of within- and between-level loadings. Although not a particular focus of this investigation, the constraint of factorial invariance in terms of L1 and L2 factor loadings is an important issue that needs further research. These between-level invariance constraints lead to measurement models which guarantee that ACH on the between level is indeed the average ACH in each cluster but becomes tenuous without such invariance constraints. With this cross-level invariance, there is a reasonably straightforward measurement model such that the latent variable is a simple decomposition of the within and between components and the between component is simply the cluster mean value. Also, as noted earlier, this invariance facilitates the interpretation in estimating the BFLPE as the difference between two regression coefficients. Fortunately, there was clear support for this cross-level invariance in this investigation in that goodness of fit indices that control for parsimony were nearly the same for models with and without such constraints. If there is no support for this sort of invariance, it might be possible to impose partial invariance in which some of the factor loadings for each latent factor are invariant over levels. Having invariance across the two levels clearly facilitates interpretations. However, we do not claim that such cross-level invariance is a necessary condition for applying multilevel contextual models with multiple indicators but only that further research into alternative solutions is needed when such invariance is not met. Interestingly, this

is not an issue for manifest models in which each of the constructs is represented by a single indicator (e.g., M1 and M2); there is an implicit cross-level invariance with all factor loadings fixed to 1.0 that provides a common metric at both levels. This is the reason, for example, results based on within- and between-group centering are mathematically equivalent, one solution being a simple algebraic transformation of the other (Raudenbush & Bryk, 2002). It is important to note that invariance constraints ensure that there is a common metric at both within and between levels and facilitates interpretation. However, the measurement might not be invariant across levels and, indeed, the whole factor structure might be entirely different across levels (e.g., Härnqvist, 1978; Zimprich, Perren, & Hornung, 2005). Clearly an important area for further research is a better understanding of the implications of violations of the assumption of cross-level invariance; what interpretations are appropriate under these circumstances and how robust interpretations are to a lack of invariance.

Effect sizes based on multilevel contextual models. Standardized parameter estimates are routinely used to summarize the results of single-level multiple regression models of manifest variables and SEM latent variable models. The standardized solutions facilitate interpretations of the results, comparison of effect sizes associated with different independent variables, and incorporation of results from different studies into meta-analyses. However, the problem of how to compute effect sizes is more complex for MLMs—particularly for latent variable models with multiple indicators—and apparently has not been resolved. Hence, we discuss alternative approaches to the computation for effect sizes for MLMs that can be applied to both manifest and latent approaches based on single-indicator or multiple-indicator constructs. We also demonstrate how standard errors of alternative effect size measures can be estimated for MLMs within the framework developed here.

Mplus currently achieves standardization for a MLM like M4 separately for each level-treating them almost as multiple (separate) groups. This is reasonable when the researcher wants to evaluate these coefficients separately for the within and between levels. However, for contextual studies, researchers need to consider coefficients between the two levels so that the default standardized coefficients are not particularly useful. It is possible to overcome this limitation by building an appropriate standardization into the model constraints. Although other researchers might choose an alternative standardization in other situations, we discuss three options here.

Tymms (2004) proposed that the effect size (ES) for continuous Level 2 predictors in MLMs, which is comparable with Cohen's d (Cohen, 1988), be calculated using the following formula:

$$\mathrm{ES1} = (2 * B * SD_{\mathrm{predictor}})/\sigma_e, \tag{8}$$

where B is the unstandardized regression coefficient in the MLM, $SD_{\text{predictor}}$ is the standard deviation of the predictor variable at L2, and σ_e is the residual standard deviation at L1. The resulting effect size describes the difference in the dependent variable between two L2 groups that differ by two standard deviations on the predictor variable (see Appendix, M8, ES1, for the operationalization of this in Mplus models considered here). Alternatively, using the same notation (Equation 8), it may be more appropriate to operationalize effect size in relation to the total variance of the L1-ASC rather than its residual, σ_e (see operationalization in M8, ES2, Appendix). Finally, it might be appropriate to standardize the BFLPE estimate with respect to the total (within + between) variance of ASC but only the between variance of ACH_B. This is appropriate because the BFLPE coefficients can be interpreted as a multiplier of ACH_B in the parameterization when using grand mean centering (see operationalization in M8, ES3, Appendix). Although these three approaches to effect size can be easily operationalized in relation to each of the models considered here, the standardized effect sizes for M4 (based on the BFLPE = −.358, see Appendix) are ES1 = −.440, ES2 = −.357, and ES3 = −.350.

The three operationalizations of effect size all have the same numerator but differ in terms of the denominator. ES1 has the disadvantage of being in relation to the residual variance of L1 ACH, which will vary substantially in terms of the other predictor variables included in the analysis. Thus, for example, in a longitudinal study in which there is a pretest achievement measure, the residual variance might be expected to be very small so that the estimated effect size would be very large relative to the corresponding estimate based on a cross-sectional study. This problem is well known in meta-analyses studies where it is recommended that effect sizes should be standardized in relation to total variance (as in ES2) rather than residual variance (as in ES1). ES3 extends this logic to include total variance from L1 and L2 and thus is the most conservative definition of effect size (although the actual difference between ES2 and ES3 is not large in this investigation).

Although a full exploration of the issues surrounding standardization of effect sizes in MLMs and the corresponding definitions of effect sizes is clearly beyond the scope of this investigation, the development of appropriate standardized parameter estimates and effect sizes is an important direction for further research that is particularly relevant to contextual models considered here. Based on our preliminary evaluation, we suggest that ES2 or ES3 should be used instead of ES1. A particular strength of the approaches taken here is that they (as well as variations) can easily be incorporated into the estimation of the models considered. This has the advantage of providing a standard error using the delta method (Raykov & Marcoulides, 2004) to test for statistical significance and construct confidence intervals around effect size estimates (Thompson, 2002).

Assumptions of causality and underlying processes. BFLPE studies—and contextual models more generally—are largely based on correlational analyses so that causal interpretations should be offered tentatively and interpreted cautiously. Here, as with all social science research, it is appropriate to hypothesize causal relations but researchers should fully interrogate support for causal hypotheses in relation to a construct validity approach (see Marsh, 2007) based on multiple indicators, multiple (mixed) methods, multiple experimental designs, and multiple timepoints as well as testing the generalizability of the results across diverse settings and measures. The evidence for construct validity includes the content, response processes by participants, internal structure in terms of consistency and factor structure, and convergent and discriminant validity in relations with other constructs—including, for example, experimental and quasi-experimental manipulations, criterion-related validity, and validity generalization to relevant and similar situations or populations. Although stronger inferences about causality are possible in longitudinal, quasi-experimental, and true experimental (with random assignment) studies, trying to "prove" causality is usually a precarious undertaking. Even in true experimental studies in applied social science disciplines, there is typically some ambiguity as the interpretation of what was actually manipulated, how it varies with different subgroups within the population, and its relevance to theory and practice. The problems of casual interpretations with contextual studies have been discussed extensively in the organizational psychology (Bliese et al., 2007) and in the social sciences more generally (e.g., Morgan & Winship, 2007).

Fortunately, there is now a growing body of BFLPE research that addresses many of these concerns (see Marsh, 2007; Marsh, Seaton, et al., 2008). Quasi-experimental, longitudinal studies based on matching designs as well as statistical controls show that ASC declines when students shift from mixed-ability schools to academically selective schools over time (based on pre-post comparisons) and in relation to students matched on academic ability who continue to attend mixed-ability schools. Extended longitudinal studies show that the BFLPE grows more negative the longer students attend a selective school and is maintained even 2 and 4 years after graduation from high school. Also, there is good support for the convergent and discriminant validity of the BFLPE as it is largely limited to academic components of self-concept and nearly unrelated to nonacademic components of self-concept and to self-esteem. Cross-national comparisons based on OECD-PISA data from 26 countries shows that the BFLPE has good cross-national generalizability. Although the "third variable" problem is always a threat to contextual studies that do not involve random assignment, Marsh, Hau, and Craven (2004; Marsh, Seaton, et al., 2008) argue that this is an unlikely counterexplanation of BFLPE results. In particular, most potential "third variables" (resources, per student expenditures, socio-economic status (SES), teacher qualifications, etc.) are positively related to L2-ACH so that

controlling for them would increase the size of the BFLPE (i.e., the negative effect of L2-ACH).

In summary, the results of any one contextual study are likely to provide limited basis of support for research hypotheses positing causal effects; support for the hypothesis must be examined in relation to a broadly conceived construct validity approach. However, models applied here appear to be important in more appropriately specifying contextual effect models and facilitating more extensive tests of the construct validity of interpretations based on such models.

ACKNOWLEDGMENTS

We thank Marcel Croon, Harvey Goldstein, David Kenny, Sophia Rabe-Hesketh, and Alexandre Morin for helpful suggestions on earlier versions of this article. Two of the coauthors (Tihomir Asparouhov and Bengt Muthén) are associated with Muthén & Muthén Inc., which distributes Mplus used to do the analyses in this investigation. Supplemental material is available from the Mplus Web site (http://www.statmodel.com/index.shtml; also see Appendix).

REFERENCES

Asparouhov, T., & Muthén, B. (2003). *Full-information maximum-likelihood estimation of general two-level latent variable models with missing data: A technical report*. Los Angeles: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2006, August). *Multilevel modeling of complex survey data.* Proceedings of the Joint Statistical Meeting, Seattle, WA. ASA section on Survey Research Methods, 2718–2726.

Asparouhov, T., & Muthén, B. (2007). *Constructing covariates in multilevel regression*. Mplus Web Notes: No. 11. Los Angeles: Muthén & Muthén.

Baumert, J., Bos, W., & Lehmann, R. (Eds.). (2000). *TIMSS/III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie—Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* [TIMSS/III: Third international mathematics and science study—students' knowledge of mathematics and science at the end of secondary education]. Opladen, Germany: Leske and Budrich.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco: Jossey-Bass.

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research Methods, 10,* 551–563.

Bovaird, J. A. (2007). Multilevel structural equation models for contextual factors. In T. D. Little, J. A. Bovaird, & N. A. Card. *Modeling contextual effects in longitudinal studies* (pp. 149–182). Mahwah, NJ: Erlbaum.

Bruner, J. (1996). A narrative model of self construction. *Psyke & Logos, 17,* 154–170.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Mahwah, NJ: Erlbaum.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulations of questions, design and analysis*. Stanford, CA: Stanford Evaluation Consortium.

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group level variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods, 12,* 45–57.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12,* 121–138.

Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press.

Goldstein, H. (2003). *Multilevel statistical models*. London: Edward Arnold.

Goldstein, H., & McDonald, R. P. (1988). A general-model for the analysis of multilevel data. *Psychometrika 53,* 455–467.

Härnqvist, K. (1978). Primary mental abilities of collective and individual levels. *Journal of Educational Psychology, 70,* 706–716.

Iverson, G. R. (1991). *Contextual analysis*. Sage University Paper Series on Quantitative Approaches in the Social Sciences, 07-081. Newbury Park, CA: Sage.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement, 13,* 171–183.

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

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65,* 457–474.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco: Jossey-Bass.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–21.

LaHuis, D. M., & Ferguson, M. W. (2009). the accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods, 12,* 418–443.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. West Sussex, UK: Wiley.

Lee, S.-Y., & Poon, W.-Y. (1998). Analysis of two-level structural equation models via EM type algorithms. *Statistica Sinica, 8,* 749–766.

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology, 30,* 263–285.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13,* 203–229.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2009). A 2 × 2 taxonomy of multilevel latent covariate models: Accuracy and bias trade-offs in full and partial error-correction models. Manuscript submitted for publication.

Marsh, H. W. (1984). Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education, 28,* 165–181.

Marsh, H. W. (1987). The big-fish-little-pond effect on ASC. *Journal of Educational Psychology, 79,* 280–295.

Marsh, H. W. (1991). The failure of high ability high schools to deliver academic benefits: The importance of ASC and educational aspirations. *American Educational Research Journal, 28,* 445–480.

Marsh, H. W. (1994). Using the National Educational Longitudinal Study of 1988 to evaluate theoretical models of self-concept: The Self-Description Questionnaire. *Journal of Educational Psychology, 86,* 439–456.

Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self concept in educational psychology—25th Vernon-Wall lecture series.* London: British Psychological Society.

Marsh, H. W., Chessor, D., Craven, R. G., & Roche, L. (1995). The effects of gifted and talented programs on ASC: The big fish strikes again. *American Educational Research Journal, 32,* 285–319.

Marsh, H. W., & Craven, R. (2002). The pivotal role of frames of reference in ASC formation: The big fish little pond effect. In F. Pajares & T. Urdan (Eds.), *Adolescence and education* (Vol. 2, pp. 83–123). Greenwich, CT: Information Age.

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163.

Marsh, H. W., & Hau, K. T. (2003). Big fish little pond effect on academic self-concept: A crosscultural (26 country) test of the negative effects of academically selective schools. *American Psychologist, 58,* 364–376.

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

Marsh, H. W., Hau, K., & Craven, R. G. (2004). The big-fish-little-pond effect stands up to scrutiny. *American Psychologist, 59,* 268–271.

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality & Social Psychology, 78,* 337–349.

Marsh, H. W., Martin, A. J., & Cheng, J. (2008). How we judge ourselves from different perspectives: Contextual influences on self-concept formation. In M. L. Maehr, S. Karabenick, & T. Urdan (Eds.), *Advances in motivation and achievement: Social psychological perspectives* (Vol. 15, pp. 315–356). New York: Elsevier.

Marsh, H. W., & O'Mara, A. (2008, July). *Have researchers underestimated the big-fish-little-pond-effect? Development of long-term total negative effects of school-average ability on diverse educational outcomes.* Paper presented at the 2008 International Congress of Psychology, Berlin, Germany.

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III (SDQ III): The construct validity of multidimensional self-concept ratings by late-adolescents. *Journal of Educational Measurement, 21,* 153–174.

Marsh, H. W., & Parker, J. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology, 47,* 213–231.

Marsh, H. W., & Rowe, K. J. (1996). The negative effects of school-average ability on academic self-concept-an application of multilevel modelling. *Australian Journal of Education, 40*(1), 65–87.

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., et al. (2008). The big-fish-little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20,* 319–350.

Marsh, H. W., Trautwein, U., Lüdtke, O., & Köller, O. (2008). Social comparison and big-fish-little-pond effects on self-concept and other self-belief constructs: Role of generalized and specific others. *Journal of Educational Psychology, 100,* 510–524.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality, 74,* 403–455.

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

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

Marsh, H. W., Wen, Z., Hau, K.-T., Little, T. D., Bovaird, J. A., & Widaman, K. F. (2007). Unconstrained structural equation models of latent interactions: Contrasting residual- and mean-centered approaches. *Structural Equation Modeling, 4,* 570–580.

Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal, 35,* 705–738.

McDonald, R. P. (1993). A general-model for 2-level data with responses missing at random. *Psychometrika, 58,* 575–585.

McDonald, R. P. (1994). The bilevel reticular action model for path-analysis with latent-variables. *Sociological Methods & Research, 22,* 399–413.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10,* 259–284.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge, UK: Cambridge University Press.

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

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354.

Muthén, B. O. (1994). Multilevel covariance structure-analysis. *Sociological Methods & Research, 22,* 376–398.

Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological methodology, 25,* 216–316.

Muthén, L. K., & Muthén, B. O. (2006–2008). *Mplus user's guide*. Los Angeles: Author.

Ntzoufras, I. (2009). *Bayesian modelling using WinBUGS.* Hoboken, NJ: Wiley.

O'Brien, R. M. (1990). Estimating the reliability of aggregate-level: Variables based on individual-level characteristics. *Sociological Methods and Research, 18,* 473–504.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69,* 167–190.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Raykov, T., & Marcoulides, G. A. (2004). Using the Delta Method for approximate interval estimation of parametric functions in covariance structure models. *Structural Equation Modeling, 11,* 659–675.

Schwanzer, A. D., Trautwein, U., Lüdtke, O., & Sydow, H. (2005). Entwicklung eines Instruments zur Erfassung des Selbstkonzepts junger Erwachsener [Development of a questionnaire on young adults' self-concept]. *Diagnostica, 51,* 183–194.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components.* New York: Wiley.

Segawa, E., Emery, S., & Curry, S. J. (2008). Extended generalized linear latent and mixed model. *Journal of Educational and Behavioral Statistics, 33,* 464–484.

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist, 55,* 5–14.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66,* 563–576.
Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
Swaminathan, H., & Rogers, H. J. (2008). Estimation procedures for hierarchical linear models. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modelling of educational data* (pp. 469–520). Charlotte, NC: Information Age.
Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31,* 25–32.
Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educational research* (pp. 55–66). London: National Foundation for Educational Research.
Zimprich, D., Perren, S., & Hornung, R. (2005). A two-level confirmatory factor analysis of a modified Rosenberg self-esteem scale. *Educational & Psychological Measurement, 65,* 465–481.

APPENDIX

MPLUS Syntax for Selected Models From the Present Investigation[a]

```
TITLE: MODEL 1 Doubly Manifest Multilevel Model
with manifest constructs and manifest aggregation;
DATA: File is bflpe.dat;
VARIABLE: Names are
idstud idsch sex szweig1
zmsc1 zmsc2 zmsc3 zmsc4
zmach1 zmach2 zmach3
zmsct zmacht
s_msc1 s_msc2 s_msc3 s_msc4 s_msct
   s_mach1 s_mach2 s_mach3 s_macht;
USEVAR ARE
  zmsct zmacht s_macht;
  cluster = idsch;
  within = zmacht ;
  between = s_macht;
  centering = groupmean(zmacht);
  missing are all (-99);
ANALYSIS: Type is twolevel ;
MODEL:
  %within%
  zmsct on zmacht (b_within);
  %between%
  zmsct on s_macht (b_betwn);
MODEL CONSTRAINT:
  new(bflpe);
  bflpe = b_betwn - b_within;
OUTPUT: sampstat stand tech1;
```

```
TITLE: Model 2 Manifest-Latent Model with manifest Single-indicators
 of each construct, but latent aggregation from L1 to L2.;
DATA: File is bflpe.dat;
VARIABLE: Names are
 idstud idsch sex szweig1
 zmsc1 zmsc2 zmsc3 zmsc4
 zmach1 zmach2 zmach3
 zmsct zmacht
 s_msc1 s_msc2 s_msc3 s_msc4 s_msct
    s_mach1 s_mach2 s_mach3 s_macht;
USEVAR ARE
  zmsct zmacht;
  cluster = idsch;
ANALYSIS: Type is twolevel ;
MODEL:
  %within%
  zmsct on zmacht (b_within);
  %between%
  zmsct on zmacht (b_betwn);
MODEL CONSTRAINT:
  new(bflpe);
  bflpe = b_betwn - b_within;
OUTPUT: sampstat stand tech1;
```

```
TITLE: model 3 Latent-manifest model with Latent constructs
based on multiple indicators, but manifest aggregation;
DATA: File is bflpe.dat;
VARIABLE: Names are
 idstud idsch sex szweig1
 zmsc1 zmsc2 zmsc3 zmsc4
 zmach1 zmach2 zmach3
 zmsct zmacht
 s_msc1 s_msc2 s_msc3 s_msc4 s_msct
    s_mach1 s_mach2 s_mach3 s_macht;
USEVAR ARE
  zmsc1 zmsc2 zmsc3
  zmsc4 zmach1 zmach2 zmach3
  s_mach1 s_mach2 s_mach3;
  cluster = idsch;
  within = zmach1 zmach2 zmach3;
  between = s_mach1 s_mach2 s_mach3;
  centering = groupmean(zmach1 zmach2 zmach3);
  missing are all (-99);
  ANALYSIS: Type is twolevel ; algorithm=em; mconv=1000;
  ANALYSIS: Type is twolevel ; algorithm=em; mconv=1000;
  MODEL:
    %within%
    MSC_W by zmsc1 (1) ;
    MSC_W by zmsc2 (2);
```

```
    MSC_W by zmsc3 (3) ;
    MSC_W by zmsc4 (4);
    MACH_W by zmach1 (5) ;
    MACH_W by zmach2 (6);
    MACH_W by zmach3 (7) ;
     MSC_W on MACH_W (b_within);
    %between%
    MSC_B by zmsc1 (1) ;
    MSC_B by zmsc2 (2);
    MSC_B by zmsc3 (3) ;
    MSC_B by zmsc4 (4);
    MACH_B by s_mach1 (5) ;
    MACH_B by s_mach2 (6);
    MACH_B by s_mach3 (7) ;
     MSC_B on MACH_B (b_between);
MODEL CONSTRAINT:
  new(bflpe);
    bflpe = b_between - b_within;
OUTPUT: sampstat stand tech1;


TITLE: Model 4 Doubly-latent MLM with latent constructs
based on multiple indicators and latent aggregation;
  mult indicators; no rand slopes;
DATA: File is bflpe.dat;
VARIABLE: Names are
 idstud idsch sex szweig1
 zmsc1 zmsc2 zmsc3 zmsc4
 zmach1 zmach2 zmach3
 zmsct zmacht
 s_msc1 s_msc2 s_msc3 s_msc4 s_msct
    s_mach1 s_mach2 s_mach3 s_macht;
USEVAR ARE
  zmsc1 zmsc2 zmsc3
  zmsc4 zmach1 zmach2 zmach3;
  cluster = idsch;
  centering = grandmean(zmach1 zmach2 zmach3);
  missing are all (-99);
ANALYSIS: Type is twolevel ; algorithm=em; mconv=1000;
  MODEL:
    %within%
    MSC_W by zmsc1 (1);
    MSC_W by zmsc2 (2);
    MSC_W by zmsc3 (3);
    MSC_W by zmsc4 (4);
    MACH_W by zmach1 (5);
    MACH_W by zmach2 (6);
    MACH_W by zmach3 (7);
    MSC_W on MACH_W (b_within);
    %between%
```

```
     MSC_B by zmsc1 (1);
     MSC_B by zmsc2 (2);
     MSC_B by zmsc3 (3);
     MSC_B by zmsc4 (4);
     MACH_B by zmach1 (5) ;
     MACH_B by zmach2 (6);
     MACH_B by zmach3 (7);
     MSC_B on MACH_B (b_betwn);
    MODEL CONSTRAINT:
    new(bflpe);
    bflpe = b_betwn - b_within;
OUTPUT: sampstat stand tech1;


    TITLE: Model 5A Doubly-latent MLM (Model 5) with random slopes
DATA: File is bflpe.dat;
VARIABLE: Names are
 idstud idsch sex szweig1
 zmsc1 zmsc2 zmsc3 zmsc4
 zmach1 zmach2 zmach3
 zmsct zmacht
 s_msc1 s_msc2 s_msc3 s_msc4 s_msct
    s_mach1 s_mach2 s_mach3 s_macht;
USEVAR ARE
  zmsc1 zmsc2 zmsc3
  zmsc4 zmach1 zmach2 zmach3;
  cluster = idsch;
  missing are all (-99);
  ANALYSIS:Type is twolevel random; algorithm=integration; integration=10;
      GHFIML=OFF;
      %within%
 MODEL:
 %within%
   MSC_W by zmsc1 (1);
   MSC_W by zmsc2 (2);
   MSC_W by zmsc3 (3);
   MSC_W by zmsc4 (4);
   MACH_W by zmach1 (5);
   MACH_W by zmach2 (6);
   MACH_W by zmach3 (7);
     s | MSC_W on MACH_W;
   %between%
    MSC_B by zmsc1 (1);
    MSC_B by zmsc2 (2);
    MSC_B by zmsc3 (3);
    MSC_B by zmsc4 (4);
    MACH_B by zmach1 (5) ;
    MACH_B by zmach2 (6);
    MACH_B by zmach3 (7);
     [s] (b_within);
```

```
    MSC_B on MACH_B (b_betwn);
  MODEL CONSTRAINT:
    new(bflpe);
    bflpe = b_betwn - b_within;
  OUTPUT: sampstat tech1; !stand
```

[a]The Mplus Web site (http://www.statmodel.com/index.shtml) contains the syntax for all 10 models and a more general description of Mplus convention for specifying the models (http://www.statmodel.com/ugexcerpts.shtml)