Reprint of:

Zumbo, B. D., & Forer, B. (2011). Testing and Measurement from a Multilevel View: Psychometrics and Validation. In James A. Bovaird, Kurt F. Geisinger, & Chad W. Buckendahl (Editors). ***High Stakes Testing in Education - Science and Practice in K-12 Settings,*** (pp.177-190). American Psychological Association Press, Washington, D.C..

11

TESTING AND MEASUREMENT FROM A MULTILEVEL VIEW: PSYCHOMETRICS AND VALIDATION

BRUNO D. ZUMBO AND BARRY FORER

A growing number of testing and assessment programs gather individual student or child measures, but by design, they do not make inferences or decisions about individual students or children but rather for an aggregate such as a school, school district, neighborhood, or state. We call such measurement practices *multilevel measurement*. In striking contrast to multilevel measurement, however, the widely used measurement and testing models (including our psychometric and validation models) are, by historical precedent, geared toward individual differences, as are the constructs they measure and related construct validation work.

Our purposes herein are to (a) introduce multilevel measurement; (b) contrast it with conventional views of measurement; and (c) discuss its implications for how one defines constructs, considers high stakes, and conducts theoretical and day-to-day work such as evaluating the measurement properties and

Bruno Zumbo is grateful to Barbara Plake, Mike Kane, and Bob Linn for initially encouraging him during the 2005–2009 National Assessment of Education Progress evaluation to look further into this multilevel problem. Also thanks to Clyde Hertzman and the Human Early Learning Partnership at the University of British Columbia for supporting our thinking about multilevel measurement. Thank you as well to Chad Buckendahl, Craig Deville, Terry Ackerman, Bob Linn, and Mike Kane for detailed feedback on an earlier draft of this chapter.

http://dx.doi.org/10.1037/12330-011
High-Stakes Testing in Education: Science and Practice in K–12 Settings, edited by J. A. Bovaird, K. F. Geisinger, and C. W. Buckendahl

inferences made from multilevel measures. Throughout we build from our recent experiences in large-scale and high-stakes multilevel measurement.

Recent experiences with two testing programs—(a) the Canadian (and now international) school-readiness assessment of kindergarten children, the Early Development Instrument (EDI; Janus & Offord, 2007), and (b) the technical working group for the evaluation of the National Assessment of Educational Progress (NAEP)—have highlighted for us our need to reassess how we approach the psychometrics and validation of multilevel measurement. Both testing programs deal with measures that one would tend to think of as focusing on individual differences for placement and more generally individual assessment uses. For example, school readiness has traditionally focused on identification of cognitive functioning and specific language and number skills with an eye toward gathering individual child measures that help school officials (e.g., teachers, school psychologists) ascertain whether the child will start school ready to learn and possibly inform educational planning for that child. Historical definitions of school readiness acknowledge individual approaches toward learning as well as the unique experiences and backgrounds of each child. Likewise, much educational testing and assessment in the domains of science and mathematics, for example, are focused on assessment of learning (summative) or even assessment for learning (formative), but in both cases the student's individual learning or knowledge is the focus.

Our central message is that quite contrary to conventional individual differences use of such tests, neither the EDI nor the NAEP is designed for or provides any feedback to individual student examinees or other stakeholders (e.g., paraprofessionals) for the purpose of providing feedback or planning for individual students. That is, like the NAEP, the EDI is not used for individual decision making but rather to inform policy and perhaps assess the impact of community-scale interventions and changes in the educational and social support system.

Instead of individual differences constructs, testing programs like the EDI or the NAEP involve what we call *multilevel constructs* that have emerged at the confluence of multilevel thinking (and ecological perspectives) with psychology, health, and social policy. For example, *school readiness* as measured by the EDI can be regarded as a construct in a time-varying multilevel network of contextual influences, and as such, psychometric studies should be conducted and EDI inferences should be validated in a way that takes into account its multilevel nature.

A multilevel construct can be defined as a phenomenon that is potentially differentially meaningful both in use and interpretation at the level of individuals and at one or more levels of aggregation. Although we focus herein on aggregate-level measures, this definition of multilevel constructs allows for measures that are used and scores that are reported only at the aggregate level

(e.g., the NAEP and some international assessments such as the Trends in International Mathematics and Science Study or the Program for International Student Assessment) as well as for measures that are used and scores that are reported at both the individual and aggregate levels (e.g., statewide educational assessments). Although all constructs reside at one level at least, an organizational setting like formal education is inherently multilevel given the natural nesting of students within classes within schools within school districts. Having to deal with multilevel issues should be assumed when studying phenomena in these multilevel settings (e.g., Klein, Dansereau, & Hall, 1994; Morgeson & Hofmann, 1999).

The essential feature is that these multilevel measures are not conventional educational achievement or psychological measures because they have been designed to provide only aggregate-level information, for example, tracking how a state is performing on a mathematics or science assessment. This aggregate-level information is in contrast to the typical use of educational and psychological measures that are used for assessment of individual differences.

Before turning to the question of measurement validation per se, it seems fitting to say a few words about how and in what way multilevel measures may be high stakes. A common feature of high-stakes testing with individual differences measures is that the test taker is directly impacted by the use and interpretation of the test scores. On the other hand, for multilevel measures, one may conclude that the stakes are actually not very high for the individual test taker—in our examples the children or students. Because of the multilevel nature of the assessment system, the testing and assessment results are designed to provide only aggregate-level information. It should be noted, however, that multilevel measurement can be, and often by the very nature of its use in shaping policy and day-to-day initiatives is, high stakes. That is, the multiple levels of the multilevel measurement system do not buffer the child or student from implications of assessment use. An example of these high-stakes results is discussed by Linn (2006, 2008) and Kane (2006) when considering the consequences of test use in policymaking and evaluation. We return to the issue of high stakes when discussing the need for evaluating the inferences made from multilevel measures (i.e., measurement validity) and the validation process of multilevel measurement.

MULTILEVEL VALIDATION

The primary question in multilevel validation concerns theoretical explanations for data variability (see Zumbo, 2007, 2009, for an explication of validity from an explanation-focused point of view). In the multilevel measurement context, this translates to addressing what constitutes the level of

theory. Clearly defining the level of theory in organizations has often proved to be problematic across many diverse multilevel settings (Dansereau, Cho, & Yammarino, 2006). In multilevel settings, the level of measurement and/or the level of statistical analysis may not be identical to the level of theory, leading to potentially spurious inferences.

Level of theory, which refers to theoretical explanations for data variability (Klein et al., 1994), has been a challenge to clearly define in many areas of multilevel research (Dansereau et al., 2006). Inferential fallacies most often occur as a result of lack of clarity in defining the level of theory in inherently multilevel settings such as school systems. However, it is absolutely necessary to strive for a theoretical basis for inferences because when the level of measurement (i.e., data) and/or the level of statistical analysis are not identical to the level of theory, a fallacy of the wrong level (Klein et al., 1994) may result. In other words, an incorrect inference may be made in which a phenomenon (e.g., an effect) is attributed to one level (e.g., schools) when it actually exists at another level (e.g., individuals).

There are two basic forms of fallacies of the wrong level. The first is the *ecological fallacy*, in which unjustified inferences are drawn at the individual level on the basis of data from some aggregation of individuals (e.g., classes, schools, states, or even countries). We illustrate by using Diez-Roux's (1998) example: A finding that countries with higher median incomes are associated with higher rates of vehicular mortality does not allow an inference that the same association holds for individuals within each country. It is entirely possible that for individuals, an inverse relationship may hold.

The second type of fallacy is the *atomistic fallacy*, in which unjustified inferences are drawn at the aggregate level on the basis of data from individuals. Bliese and Halverson (1996), for example, showed that the (negative) correlation between work hours and well-being is much smaller at the individual level than at the workgroup level. Therefore, any inferences made at the group level on the basis of the individual-level results would be incorrect, missing a relationship that emerges only at a higher level.

The atomistic fallacy is particularly germane in the context of measures like the EDI and the NAEP, which have been designed for interpretation only at a group level. Indeed, any interventions based on EDI or NAEP results are targeted at groups rather than at individual children. The effectiveness of these interventions should therefore be based on group-level data to avoid making an atomistic fallacy (Bliese, 2000). To emphasize, to avoid inferential fallacies, multilevel researchers need to match the level of data with the level at which inferences are desired.

Although evidence is needed to support such assertions, it may be argued that with multilevel measures like the EDI or the science and mathematics

assessments of the NAEP, the primary dimension being assessed at the individual level remains the same across some levels, but secondary dimensions may arise at higher levels. That is, a secondary dimension of teacher or classroom effect or perhaps neighborhood characteristics, curricular differences, or opportunity-to-learn differences emerge across states. In short, any inferences from the individual level may not hold in the same way at higher (or lower) levels of aggregation.

At the very least, systematic and coherent validation evidence needs to be assembled to support the inferences at the various levels. Furthermore, the level of validation evidence needs to be in line with the level of inferences. Therefore, individual-level validity evidence (which is what is traditionally involved in validation research, such as criterion validity at the child level) does not provide sufficient validity evidence for inferences at higher levels in the system and may actually be misleading because it may miss invalidity at the aggregate level.

In short, the need for multilevel validation arises when one has a multilevel construct; measurement (or assessment) that occurs at the individual level and individual responses are aggregated to make inferences at a higher level. Historically, multilevel constructs have not been a widespread issue in measurement and validation because traditional views of measurement and assessment have been immersed in and emerged from an individual-differences psychological school of thought, such as Cronbach and Meehl (1955) in psychological measurement. Individual-differences researchers investigate the ways in which individual people differ in, for example, their cognitions, behavior, attitudes, aptitudes, emotions, or even physiologically. The tests and measures used at the individual level are developed for the purpose of investigating ways in which individual people differ. We do not discuss these measures herein because there are many examples and a long history of individual-differences measures of school readiness and of science or mathematics knowledge and achievement.

To this point, our central messages and their implications have been that multilevel constructs are different in purpose and scope from individual-differences constructs, although they still potentially carry high stakes for the individual test taker. Likewise, multilevel constructs necessitate multilevel measures. Multilevel measurement and testing arise when one has a multilevel construct, that is, an individual-level measure (or assessment) that one aggregates to make inferences at a higher level. Historically, multilevel constructs have not been a widespread issue in measurement and validation because testing and measurement have been immersed in and emerged from individual differences. Implied in our views is that applying only traditional individual-differences psychometric methods (e.g., correlation with another

child school-readiness measure) and/or most cognitive assessment approaches is insufficient to gather evidence for the support of multilevel validation inferences using assessments like the EDI or the NAEP. In fact, individual-differences methods are susceptible to the cross-level inferential fallacies such as the ecological fallacy or atomistic fallacy. Given the move to increase the use of assessment results in the formulation of policy and the shift in educational and psychological theorizing toward ecological views of our phenomenon, we fully expect to see more multilevel constructs in the coming years.

IMPLICATIONS FOR THEORETICAL AND OPERATIONAL WORK

We now turn to the implications for multilevel constructs and aggregate uses of test data for theoretical and operational (day-to-day research) work. As a starting point, multilevel psychometric research might address a number of fairly generic multilevel research questions. Some of these questions may include: (a) Does the aggregate score reflect differences between measurement units at the aggregate level, such as neighborhood differences in school readiness? (b) To what extent might other important constructs be measured unintentionally at the aggregate level that are not meant to be included in an assessment of school readiness at the aggregate level, such as classroom or teacher effects, neighborhood effects, or regional effects? or (c) When assessment data are considered at the state level, how much of the variation is attributable to state-to-state differences relative to student-to-student differences?

We now, however, turn to the fairly traditional measurement issues and questions of reliability, validity, and use of assessment data (as well as the reporting results) to see how they turn out to be influenced by multilevel constructs.

Reliability of Measurement

To begin, it should be noted that even the notion of quantifying measurement error (through reliability of measurement indices) becomes very complex, with several subtle issues, such as defining the unit of reference in terms of the domain scores. Of course, test–retest evidence with the aggregate unit of analysis could be used, but that also makes certain assumptions about the sources of error variance. In terms of multilevel measurement, most of the systematic work done to date involves the reliability of measurement.

There already exists a nice motivating context for multilevel reliability of measurement in terms of the work done on the reliability of class means (e.g., teaching evaluation data or student classroom assessment results). When classes are the units of analyses, estimates of the reliability of class means are needed. If one uses classical test theory, it is difficult to treat this problem adequately;

however, generalizability theory, which is by design a multilevel measurement model, provides a framework for dealing with the problem. However, if one builds on existing literature from work on generalizability of class means, it becomes apparent that the reliability of student-level assessment can sometimes be greater and other times smaller than the aggregate level. Although summarizing this literature in detail is beyond the scope of this chapter, interested readers should see Brennan, Yin, and Kane (2003); Gillmore, Kane, and Naccarato (1978); Gillmore, Kane, and Smith (1984); Kane and Brennan (1977); Kane, Gillmore, and Crooks (1976); O'Brien (1990); and Yin and Brennan (2002) for a thorough treatment of the subject.

Clearly, a central issue in multilevel measurement reliability is that individual-level (person-level) reliability data cannot be counted on to make measurement reliability claims at the aggregate level. As is evident from the pioneering work of Kane, Brennan, Gillmore, and others, the level of reliability data must be matched with the level of data interpretation or the wrong conclusion about measurement reliability may result.

We now turn to multilevel validity and the process of validation.

Validity and the Process of Validation

The appropriate process of measurement validation for multilevel constructs is a pressing issue and one that has not been explicitly dealt with in the educational measurement literature. As we noted previously, the key is in the interpretation and use of test results at the aggregate level.

What is clear is that several validity theorists and practitioners have talked at the edges of the multilevel validity (and multilevel measurement) issue, but no one seems to have taken it on at the forefront. Of course, the general theories of validity per se are writ so large that one could easily directly apply some of them (e.g., Cronbach & Meehl, 1955; Kane, 2006; Messick, 1989; Zumbo, 2007, 2009). The issue here, however, comes to (a) what is meant by *validity* and (b) the many cases in which the explanatory power of the historically dominant views of validity in the assessment field is in individual differences, which in a sense diminishes the importance of aggregation and the effect of contextual variables that exist at various levels of the multilevel system. The importance of aggregation and the influence of variables at various levels of the multilevel system are limitations of taking on some of the traditional approaches to guide validity of the multilevel assessment. The interested reader should see Kane (2006) and Zumbo (2007, 2009) for a review of contemporary thinking in validity as well as Linn (2006, 2008) for a discussion of validating school quality inferences from student assessments.

One certainty, though, is that measurement programs need to have an articulated, coherent, and a consistent validation plan that amasses the

theoretical and empirical evidence that supports the inferences being made with the test or measure at the various levels.

Zumbo's Draper–Lindly–DeFinetti framework (Zumbo, 2007) puts importance on the sampling units (the respondents) and their characteristics, something that is not highlighted enough in the conventional discussions of psychometrics and validity evidence. Most, but not all, validation studies in the research literature give little time to the exchangeability of the sampled and unsampled units. In addition, there is little discussion in the psychometric literature of matters of complex sampling. The complex multilevel assessment data now being collected by many government, health, and social science organizations around the world have increasingly complex structures, precipitating a need for ways of incorporating these complex sampling designs into psychometric models. It is worth noting that the Draper–Lindly–DeFinetti framework framework shines a spotlight on the person-sampling aspect of measurement, which has mostly held a secondary place to item or domain sampling in psychometrics.

Whenever data are aggregated over one or more levels (e.g., schools, neighborhoods, states), the procedure must be justified in terms of establishing an alignment between the nature of the construct, its measurement, and its analysis vis-à-vis other constructs of interest. In Table 11.1, we build on Chen, Mathieu, and Bliese (2004a, 2004b) for a series of adapted step-by-step procedures for conducting multilevel construct validation.

The first step in the Chen et al. (2004a, 2004b) framework deals with the theoretical issues of construct definition, such as the construct's domain boundaries and dimensionality. The purpose of this step is to establish the

TABLE 11.1
Proposed Steps for Multilevel Construct Validation

Step	Description
1	Establish construct definition at each level and the nature of the construct at aggregate level(s).
2	Specify the nature and structure of the aggregate construct (i.e., select an appropriate composition model).
3	Gather evidence appropriate to the psychometric properties across levels and multilevel latent variable (i.e., factor analysis and item response) modeling.
4	For construct variability within and between units ■ Ensure there is sufficient variability within and between units (i.e., at lower and higher levels). ■ For some aggregate-level measures, intermember reliability (intraclass correlation coefficients) can provide relevant evidence.

extent to which the meaning of a construct does or does not differ across levels.

The second step in the framework is articulating the nature of the aggregate construct. There are two basic categories of aggregate measures: (a) global measures that describe the group as a whole and (b) measures that summarize a collection of lower level (usually individual) scores (Hofmann & Jones, 2004). The EDI and the NAEP are examples of an aggregate construct of the latter category. When aggregating scores, there are at least six compositional models from which to choose (see Chen et al., 2004a). The appropriateness of a particular compositional model depends on both one's multilevel theoretical expectations and observed patterns of within-group and between-groups variation.

The first compositional model is the *selected-score* model, in which the score of one individual characterizes the group-level construct. The second is the *summary index* model, in which the group construct is based on a statistical summary (typically the mean or sum) of individual scores in a group. This is the compositional model most often used to create aggregate-level EDI or NAEP scores. The third is the *consensus* model, in which group-level constructs capture within-group agreement based on items that refer to the individual (e.g., individuals asked to rate their own teaching effectiveness). The fourth model is the *referent-shift consensus* model, which differs from the consensus model only in that it captures within-group agreement based on items that refer to the group (e.g., individuals asked to rate their department's teaching effectiveness). The fifth model is the *dispersion* model, which focuses on within-group variability and is most often expressed in the aggregate in terms of group diversity (e.g., heterogeneity of teaching styles among department members). Chen et al.'s (2004a) sixth and final model is the *aggregate properties* model, in which group constructs are directly measured at the group level (e.g., asking a school principal to rate staff effectiveness).

The third step in the construct validation process is to gather evidence appropriate to the nature of the construct and the composition model at the aggregate level. Depending on the model, this involves considering within-group agreement on item scores, factor structure across levels, and reliability of item scores. Within-group agreement, for instance, is particularly relevant for compositional models based on consensus of individuals. With regard to factor structure, the amount of expected similarity across levels should be theory driven (Chen et al., 2004a). Finally, depending on the compositional model, reliability of the item scores at the group level can be calculated quite differently because of different assumptions about systematic and error variance.

The fourth step in the multilevel construct validation process is an analysis of the relative amounts of within-group and between-groups variation,

which provides empirical guidance about appropriate levels of aggregation. Bliese (2000) and Chen et al. (2004a) discussed three different measures that can help assess whether data collected from individuals have group-level properties. First, the level of nonindependence in data can be measured using an intraclass correlation coefficient, or ICC(1), which represents the proportion of individual variance that is influenced by or depends on group membership. A second important aspect of potential aggregation is between-groups reliability, or ICC(2), which indexes the reliability of differences between group means by measuring the proportional consistency of variance across groups (Bliese, 2000). The third measure is within-group reliability, which is the degree to which group means can be reliably estimated even when group size is relatively small. The fourth aspect of group properties is within-group agreement, most commonly measured using the r_{wg} statistic.

Within-and-between analysis (Dansereau & Yammarino, 2000) is an alternative multilevel validation technique that compares patterns of within-group and between-groups variability to determine appropriate levels of aggregation. Kim (2004) pointed out that the ICC approach to comparing between-groups and within-group variance suggested by Bliese (2000) and Chen et al. (2004a) works well for constructs based on between-groups variance (i.e., analysis of variance model) but not as well as the within-and-between analysis approach for constructs based on within-group variance.

Of particular methodological importance are the multilevel latent variable modeling strategies for specification of both between-groups and within-group latent variable models by Grilli and Rampichini (2007); Muthén (1994); Rabe-Hesketh, Skrondal, and Pickles (2004); and Rijmen, Tuerlinckx, De Boerck, and Kuppens (2003). In particular, those interested in predictive validity evidence across levels of aggregation (e.g., how student-level data can be used for predictive validity studies when making state-level comparisons) should consult Croon and van Veldhoven (2007). This research illustrates (a) how regression models that are conducted at the aggregated level (a common practice when conducting criterion validity studies in multilevel measurement settings) result in biased parameter estimates and, hence, for our purposes, incorrect validity conclusions and (b) a latent variable multilevel model to correctly perform these analyses.

Use of Multilevel Assessment Data and Reporting Results

We have only some minor observations on the use of multilevel assessment data and the reporting of results. In particular, we want to highlight how the use of assessment and the reporting of results tie into and highlight the high-stakes notion of these multilevel assessments. First, it is important to note that

Kane's (2006) review chapter on validity deals with a related and central issue of consequential aspects of validity and how this connects to accountability in, for example, program evaluation. Second, Linn (2008) put the matter of the need for what we would call multilevel construct validity most succinctly; when discussing the validity of school quality inferences from student assessments, he wrote the following:

> The use of student assessment results to identify schools that need improvement and are therefore subject to various types of corrective actions or sanctions while other schools are identified as making adequate yearly progress rests on an implicit assumption that the observed school-to-school differences in student achievement are due to differences in school quality. . . . The validity of the school quality inference needs to be evaluated. (Linn, 2008, p. 12)

From our point of view, Linn (2008) was challenging the field to think about the potential errors in inference that can be made across levels of data, that is, what was previously referred to as *ecological* or *atomistic fallacies* of data inferences. In fact, the potential error in inference across levels of data further opens up Messick's as well as Cronbach's notions of the consequential considerations of test use and reporting and how these play into the validity of the measurement inferences. In this light, one needs to think about the cross-level consequences and the eventual trickle-down of the high stakes resulting from the use and reporting of multilevel assessment data.

SUMMARY

In summary, applying traditional individual-differences validation methods (e.g., correlation with another individual-differences measure) is insufficient to gather evidence to support multilevel validation inferences. In fact, individual-differences validation methods are susceptible to the cross-level inferential fallacies, such as the ecological or atomistic fallacies. Multilevel measurement and multilevel construct validation involves steps beyond the validation of single-level constructs. An important point to keep in mind is that even highly isomorphic multilevel constructs can have similar and distinct antecedents, correlates, and outcomes across levels. Assuming that they are only similar can lead to cross-level inferential fallacies from individual-level data. From the vantage point of the successes achieved by the NAEP and the EDI, the move to the increased policy usage of assessment results, and the shift in educational and psychological theorizing toward ecological views of our phenomenon, we fully expect to see more multilevel constructs in the coming years.

REFERENCES

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

Bliese, P. D., & Halverson, R. R. (1996). Individual and nomothetic models of job stress: An examination of work hours, cohesion, and well-being. *Journal of Applied Social Psychology, 26,* 1171–1189. doi:10.1111/j.1559-1816.1996.tb02291.x

Brennan, R. L., Yin, P., & Kane, M. T. (2003). Methodology for examining the reliability of group mean difference scores. *Journal of Educational Measurement, 40,* 207–230. doi:10.1111/j.1745-3984.2003.tb01105.x

Chen, G., Mathieu, J. E., & Bliese, P. D. (2004a). A framework for conducting multilevel construct validation. In F. J. Yammarino & F. Dansereau (Eds.), *Research in multilevel issues: Multilevel issues in organizational behavior and processes* (Vol. 3, pp. 273–303). Oxford, England: Elsevier. doi:10.1016/S1475-9144(04)03013-9

Chen, G., Mathieu, J. E., & Bliese, P. D. (2004b). Validating frogs and ponds in multilevel contexts: Some afterthoughts. In F. J. Yammarino & F. Dansereau (Eds.), *Research in multilevel issues: Multilevel issues in organizational behavior and processes* (Vol. 3, pp. 335–343). Oxford, England: Elsevier. doi:10.1016/S1475-9144(04)03016-4

Cronbach, L. J., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. doi:10.1037/h0040957

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods, 12,* 45–57. doi:10.1037/1082-989X.12.1.45

Dansereau, F., Cho, J., & Yammarino, F. J. (2006). Avoiding the "fallacy of the wrong level." *Group & Organization Management, 31,* 536–577. doi:10.1177/1059601106291131

Dansereau, F., & Yammarino, F. J. (2000). Within and between analysis: The varient paradigm as an underlying approach to theory building and testing. In K. J. Klein & S. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 425–466). San Francisco, CA: Jossey-Bass.

Diez-Roux, A. V. (1998). Bringing context back into epidemiology: Variables and fallacies in multilevel analysis. *American Journal of Public Health, 88,* 216–222. doi:10.2105/AJPH.88.2.216

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement, 15,* 1–13. doi:10.1111/j.1745-3984.1978.tb00051.x

Gillmore, G. M., Kane, M. T., & Smith, P. L. (1983). The dependability of student evaluations of teaching effectiveness: Matching conclusions to designs. *Educational and Psychological Measurement, 43,* 1015–1018.

Grilli, L., & Rampichini, C. (2007). Multilevel factor models for ordinal variables. *Structural Equation Modeling, 14,* 1–25. doi:10.1207/s15328007sem1401_1

Hofmann, D. A., & Jones, L. M. (2004). Some foundational and guiding questions for multi-level construct validation. In F. J. Yammarino & F. Dansereau (Eds.), *Multi-level issues in organizational behavior and processes* (pp. 305–315). Amsterdam, The Netherlands: Elsevier. doi:10.1016/S1475-9144(04)03014-0

Janus, M., & Offord, D. R. (2007). Development and psychometric properties of the Early Development Instrument (EDI): A measure of children's school readiness. *Canadian Journal of Behavioural Science, 39,* 1–22. doi:10.1037/cjbs2007001

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 47,* 267–292.

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. doi:10.1111/j.1745-3984.1976.tb00009.x

Kim, K. (2004). An additional view of conducting multi-level construct validation. In F. J. Yammarino & F. Dansereau (Eds.), *Multi-level issues in organizational behaviour and processes* (pp. 317–333). Amsterdam, The Netherlands: Elsevier. doi:10.1016/S1475-9144(04)03015-2

Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review, 19,* 195–229. doi:10.2307/258703

Linn, R. L. (2006). Validity of inferences from test-based educational accountability systems. *Journal of Personnel Evaluation in Education, 19,* 5–15. doi:10.1007/s11092-007-9027-6

Linn, R. L. (2008). *Validation of uses and interpretations of state assessments.* Washington, DC: Council of Chief State School Officers.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Morgeson, F. P., & Hofmann, D. A. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Review, 24,* 249–265. doi:10.2307/259081

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22,* 376–398. doi:10.1177/0049124194022003006

O'Brien, R. M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. *Sociological Methods & Research, 18,* 473–504. doi:10.1177/0049124190018004004

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika, 69,* 167–190. doi:10.1007/BF02295939

Rijmen, F., Tuerlinckx, F., De Boerck, R., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185–205. doi:10.1037/1082-989X.8.2.185

Yin, P., & Brennan, R. L. (2002). An investigation of difference scores for a grade-level testing program. *International Journal of Testing, 2,* 83–105. doi:10.1207/S15327574IJT0202_1

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 45–79). Amsterdam, The Netherlands: Elsevier Science.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65–82). Charlotte, NC: Information Age.