VALIDATION OF MULTILEVEL CONSTRUCTS: METHODS AND EMPIRICAL FINDINGS FOR THE EARLY DEVELOPMENT INSTRUMENT

by

BARRY ALLAN FORER

B.Sc., The University of Toronto, 1979
M.A., The University of Victoria, 1987

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2009

Abstract

A growing number of assessment, testing and evaluation programs gather individual measures but, by design, do not make inferences or decisions about individuals but rather for an aggregate such as a school, school district, neighbourhood, or province. In light of this, a multilevel construct can be defined as a phenomenon that is potentially meaningful both at the level of individuals and at one or more levels of aggregation. The purposes of this dissertation are to highlight the foundations of multilevel construct validation, describe two methodological approaches and associated analytic techniques, and then apply these approaches and techniques to the multilevel construct validation of a widely used school readiness measure called the Early Development Instrument (EDI). Validation evidence is presented regarding the multilevel covariance structure of the EDI, the appropriateness of aggregation to classroom and neighbourhood levels, and the effects of teacher and classroom characteristics on these structural patterns. To appropriately assess the multilevel factor structure of the categorical EDI items, a new fit index was created. A good-fitting unidimensional model was found for each scale at the level of individual students, with no notable improvements after taking clustering into account. However, at the class and neighbourhood levels of aggregation, the physical and emotional EDI scales did not show essential unidimensionality. Teacher and/or classroom influences accounted for between 19% and 25% of the total variance. EDI emotional scores were higher for teachers with graduate training, while communications scores were higher for younger teachers. Teachers tended to rate students more absolutely, rather than relative to other children in the class, when class size was small. These results are discussed in the context of the theoretical framework of the EDI, with suggestions for future validation work.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I owe a great debt of gratitude to the Human Early Learning Partnership at UBC. As an organization, it provided support to my dissertation research through the Thesis Grant, and more importantly, is a very stimulating place to work. I appreciate working with such congenial, motivated people with similar yet complementary interests.

I am very thankful of my committee – Professor Bruno Zumbo, Professor Hillel Goelman, and Professor Jennifer Shapka. You have all been so supportive of my research, and so helpful in making it dance. In particular, Professor Goelman has been an inspiration for me for the 20 years that we have known one another. I would not have even started this PhD without his active encouragement.

My supervisor, Professor Zumbo, has proven his wisdom and patience time and time again. Even though he is younger than me, I continually aspire to learn from him, and look forward to lots of collaboration in the future.

My parents have stood by me throughout – watchful, hopeful, and ready to celebrate.

My wife Jan, my significant everything – most of all, this dissertation is for you. Thank you for keeping the faith.

Speaking of which, thanks to the whole world for their patience and optimism. Who knew?

Chapter 1

Introduction

In education, our widely-used measurement and testing models (including our psychometric and validation models) are, by historical precedent, geared to individual differences, as are our constructs and construct validation work. However, there is a growing number of assessment, testing and evaluation programs (e.g., B.C. Foundation Skills Assessment [FSA], National Assessment of Educational Progress [NAEP]) in which one gathers individual – level measures, but inferences are limited, by design, to aggregate entities such as schools, school districts, or countries (Zumbo & Forer, in press). In such situations, multilevel approaches are necessary to avoid "the fallacy of the wrong level" (Dansereau & Yammarino, 2006). The broad aims of this dissertation are to describe multilevel construct validation, explore its foundations in the organizational psychology literature, and apply, without loss of generality, its methodologies in the context of a widely-used school readiness measure.

A consensus on the definition for the construct of *school readiness* continues to be elusive. Graue (2006) notes that it is necessarily a relational construct, and has something to do with skills and dispositions associated with success in school. Current definitions typically share the perspective of Bronfenbrenner's (1977) ecological theory of development, which would situate school readiness within a time-varying multilevel network of contextual influences, such as peers, school, home, and neighbourhood (Rimm-Kaufmann & Pianta, 2000). Given this perspective, any school readiness measure should be validated in a way that takes into account this underlying multilevel nature. This is true whether one's theory of school readiness focuses on individual differences, school differences, or population-level differences. In all cases, the validation process involves measures at more than one level, and thus requires care in reconciling levels of measurement, analysis, and theory (Klein, Dansereau, & Hall, 1994).

Whatever the level of theory, school readiness measures most typically come from individual children's scores, which are then aggregated in ways consistent with theoretical expectations. However, other conceptualizations of school readiness are also plausible, such as those based on schools being ready for children, which implies measurement at the school level (Snow, 2006).

The two purposes of this dissertation are to: 1) articulate and develop two multilevel construct validation frameworks adapted from the organizational psychology literature (Chen, Mathieu, & Bliese, 2004a; Dansereau, Alutto, & Yammarino, 1984) and 2) demonstrate how they can be applied in the context of one measure of school readiness, the Early Development Instrument (EDI; Janus & Offord, 2007). The EDI is an example of a school readiness measure that is designed for interpretation at a population level only, even though children are scored individually by their kindergarten teacher. A detailed description of the EDI follows in the literature review. Recently, Guhn and Goelman (2009) conducted an analysis of the EDI's conceptual and philosophical basis.

These purposes will be achieved in five chapters. Chapter 1 continues with a literature review, consisting of the following sections:

1.1 Validation of multilevel constructs

1.2 Two different multilevel validation frameworks

1.3 School readiness as a population-based multilevel construct

1.4 The EDI – description, level of theory, and summary of validation evidence.

1.5 Teacher effects

1.6 Three research questions

Chapter 2 provides details of the study sample and descriptive summaries of EDI scores at various levels of aggregation. In Chapter 3, the first research question is answered, starting with relevant background information, followed by the methods employed, and ending with the

results obtained. Chapter 4 follows the same pattern of background information, methods, and results, this time for research questions 2 and 3. These questions are considered together because they both are concerned with teacher and classroom effects. Finally, the results for all three research questions are discussed, including limitations and future directions, in Chapter 5.

Literature Review

1.1 Validation of Multilevel Constructs

A multilevel construct can be defined as a phenomenon that is potentially meaningful both at the level of individuals and at one or more levels of aggregation. While all constructs reside at one level at least, constructs in an organizational setting like formal education are inherently multilevel, given the natural nesting of students within classes within schools within school districts and neighbourhoods. A multilevel validation approach should be assumed when studying phenomena in these organizational settings (Keating, 2007; Klein et al., 1994).

The approach to multilevel validation in this dissertation is based on the modern unitary view (e.g., Zumbo, 2007a, 2009) of construct validity that originated with Cronbach and Meehl (1955). Construct validation is defined as the process by which researchers provide ongoing evidence to establish the range of appropriate inferences that can be made from our observed scores to our theoretical expectations (conceptualization) for a particular construct, taking into account all potential ethical and social influences (Messick, 1995). This unitary view is in contrast to the mid-20th century view (American Psychological Association, 1954) of four distinct types of validities: construct, concurrent, predictive, and content, and the adjustment in the 1980s to a three-category taxonomy, combining concurrent and predictive validities into criterion-related validity.

This unitary approach implies that inferences are validated rather than scores *per se*, so that the inferential "leap" is from the scores to the construct, rather than to the intermediary latent variable. In addition, this approach implies that it is insufficient to rely on weaker forms of validity (i.e., correlations with other variables) that are commonly in use. Instead, validity means having a model that *explains* the variation in observed scores as well as covariation with theoretically and/or practically relevant explanatory variables (Zumbo, 2007a, 2009); in other words, variables within the construct's nomological network (Cronbach & Meehl, 1955). As Zumbo (2007a) has pointed out, measurement inferences should also assign equal importance to the sampling of units (e.g., people) and items, as exchangeability is important for both of these dimensions.

In a multilevel validity context, these explanatory models must include an answer about the level of one's theory (i.e., the level of aggregation where inferences are desired). Knowing one's level of theory is important because it is not necessarily identical to either the level of measurement (the actual source of the data) or the level of statistical analysis (how data are treated) (Klein et al., 1994). Therefore, a comprehensive construct validation framework is needed, given that constructs can be aggregated in different ways, with appropriate validation evidence required for each (Chen et al., 2004a).

When inferences are made at the levels of measurement or statistical analysis rather than the level of theory, there is the risk of spurious conclusions, either due to an ecological or an atomistic fallacy. An ecological fallacy (Robinson, 1950) refers to theoretically or empirically unjustified inferences made about individuals based on group-level results, while an atomistic fallacy (Diez-Roux, 1998) refers to unjustified inferences about groups based on individual results. Therefore, as Zumbo and Forer (in press) state, the level of validation evidence needs to

be in line with the level of inferences and that validity evidence at the individual level is not only insufficient to support inferences at an aggregate level such as schools, but it may be misleading.

1.2 Two Different Multilevel Validation Frameworks

Organizational psychology is a field where methodologists have long been at the forefront of theory and techniques for multilevel construct validation. The application of these techniques to the context of educational research has been relatively sparse. Most educational researchers are content to describe in detail how scores are aggregated, but do not provide any justification for when and why inferences are appropriate at the group or individual level (Griffith, 2002).

In this dissertation, two competing methodological approaches to multilevel construct validation from organizational psychology are described and compared. Both frameworks are then applied to the validation of the EDI as a multilevel construct.

1.2.1 Chen et al. (2004a) Framework

Chen et al. (2004a) have described a set of step-by-step procedures for conducting multilevel construct validation. There are five necessary steps in their validation framework:

1. define the construct across levels of analysis;
2. articulate the nature of the aggregate construct;
3. determine the psychometric properties of the construct across levels of analysis;
4. ensure that there is construct variability across levels of analysis;
5. examine the function of the construct across levels of analysis.

The first step is to define the construct at each level of analysis. Constructs are defined by establishing domain boundaries and dimensionality, and deciding whether the construct is

formative or reflective in nature (Bollen & Lennox, 1991; Zumbo, 2007a). The purpose of this step is to establish the extent to which the meaning of a construct differs across levels, if at all (Chen, Mathieu, & Bliese, 2004b). This is a critical step because it is necessary to understand how a construct differs across levels before the next steps in the framework can be addressed (Hofmann & Jones, 2004).

The second step is to specify the nature and structure of the aggregate construct. To increase conceptual precision for developing and evaluating constructs at different levels of analysis, researchers (e.g., Chan, 1998, Kozlowski & Klein, 2000) have created typologies for multilevel constructs based on *composition models*, which specify various possible relationships when aggregating constructs to different levels. There are a variety of compositional models to choose from, depending on theory, the purpose of the research, and practical considerations. A compositional model must be explicitly identified before gathering validation evidence, as the models each differ in their expected psychometric properties.

Chan's (1998) typology has five composition models: 1) *additive*, where the aggregate construct reflects the sum or average score of the group; 2) *direct consensus*, where the aggregate construct reflects within-group agreement among group members on items referring to the member (e.g., individuals asked to rate their own level of cooperation in their team); 3) *referent-shift consensus*, where the aggregate construct reflects within-group agreement among group members on items referring to the group (e.g., individuals asked to rate their team's level of cooperation); 4) *dispersion*, where the aggregate construct reflects the variance or diversity of group members (e.g., heterogeneity of personality types in a team); and 5) *process*, where the focus is on analogous processes (rather than attributes or outcomes) across levels. Kozlowski and Klein (2000) created what they called a *typology of emergence*, in which aggregate constructs are

classified on a spectrum from pure isomorphism (e.g., synchronized swimmers) to pure compilation (where diverse contributions of members cause a new construct to emerge).

Chen et al.'s (2004a) typology slightly expands upon that of Chan (1998), with six compositional models. The first is the *selected-score* model, where the score of one individual characterizes the group level construct. The second is the *summary index* model, which is the same as Chan's (1998) additive model. This model has been used often to create aggregate level EDI scores (e.g., Janus, Walsh, Viveiros, Duku, & Offord, 2003; LaPointe, Ford & Zumbo, 2007). Chen et al. (2004a) also includes a *consensus* model, a *referent-shift consensus* model, and a *dispersion* model. Their sixth and final model is the *aggregate properties* model, where group constructs are measured at the group level directly (e.g., asking a school principal to rate staff effectiveness).

The third step in conducting multilevel construct validation is to gather validation evidence that is specific to the nature of the construct and the composition model at the aggregate level. The psychometric requirements in terms of factor structure, within-group agreement, and reliability depend on the multilevel constructs involved. This will be discussed in more detail with respect to the EDI in Chapter 3.

The fourth step in multilevel construct validation is making sure that there is appropriate within-group and between-group variability for the construct. This information "can have important implications for inferences drawn with respect to the internal structure and external function of the multilevel construct" (Chen et al., 2004a, p. 290). A variety of within-group reliability measures can be applied at this step, as well as other techniques that compare within- and between-group variability.

Chen et al.'s (2004a) fifth and final step in multilevel construct validation is testing the function of the construct across levels. This entails testing theoretical relationships with other

constructs in the nomological network at different levels of analysis. Hofmann and Jones (2004) also emphasize that functional similarities, and not just structural similarities, are a necessary component in the assessment of the psychometric properties of a multilevel construct.

1.2.2 Dansereau, Alutto, & Yammarino (1984) Framework

An alternate approach to multilevel construct validation is called Within and Between Analysis (WABA), developed by Dansereau, Alutto, and Yammerino (1984). According to this approach, when researchers specify what they intend to describe and/or explain (i.e., the level of theory or level of inference), this entails making assumptions about the patterns of between- and within-group variability for the constructs involved. There are three idealized levels of theory. Groups can be composed of homogeneous members, independent members, or interdependent members (Klein et al., 1994). Dansereau and Yammarino (2000) refer to these three situations as *wholes*, *equivocal*, and *parts*, respectively. With *homogeneous* members (wholes view), a single value is sufficient to describe the group. The assumed pattern for this level of theory is that there is only between-group variation/covariation for a construct or the relationship between constructs. With *independent* members (equivocal view), all of the variation/covariation is assumed to be interindividual (i.e., no distinction between within-group and between-group variation), and thus due to individual variation only. Finally, with *heterogeneous* members (parts view), there is individual variation, but only relative to the group context. In this assumed level of theory, variation/covariation is predominantly within-group. In summary, the wholes view stresses the exchangeability of individual members, the equivocal view points to a lack of group influence, and the parts view implies that individual scores have only relative meaning.

By examining variability patterns, WABA analyses (Dansereau & Yammarino, 2000) determine the appropriateness of aggregating data to a particular level. WABA inferences about

the prevailing ideal pattern are made only after considering both practical significance as well as statistical significance. Practical significance is assessed using a statistic called the E-ratio (described below), while statistical significance is assessed using the more familiar F-ratio. For WABA analyses, the inference implied by the E-ratio is accepted only if confirmed statistically using the corresponding one-way ANOVA F-ratio (Dansereau & Yammarino, 2000). One important wrinkle in the use of the F-ratio in WABA is that the numerator and denominator in the ratio depend on the inference implied by the E-ratio. When a "wholes" inference is implied, the F-ratio is calculated as the typical MS(between)/MS(within). When a "parts" inference is implied, the inverse ratio is calculated.

The E-ratio, used to assess practical significance, is calculated as the between-group eta correlation divided by the within-group eta correlation. The between-group eta correlation is a sort of effect size for variation between groups (e.g., classes), and the within-group eta correlation is the corresponding effect size for variation within groups. Each has a range from 0 to 1. When within-group and between-group variation are equal, the E-ratio is equal to 1. Dansereau and Yammarino (2000) demonstrate that the E-ratio is mathematically equivalent to Cohen's (1988) *f* effect size statistic.

The E-ratio ranges from a low of 0 (no variability between groups) to infinity (no variability within groups), assuming non-zero total variation. Just like the F-ratio, one or more predetermined critical values of the E-ratio need to be established before making inferences. Dansereau et al. (1984) have established two sets of geometrically-determined critical values, the 15^{o} test and the 30^{o} test. The former divides the overall variance into approximately equal intervals for the purposes of the three possible inferences, while the latter uses equal intervals in terms of angles (90^{o} divided by 3). The 30^{o} test is more stringent, requiring between-group or

within-group variation to exceed 75% of the total variation before an inference of wholes or parts, respectively, may be made.

The WABA approach to multilevel construct validation can be applied to four inferential purposes: single level analysis (SLA), multiple level analysis (MLA), multiple variable analysis (MVA), and multiple relationship analysis (MRA). The purpose of SLA is to detect whether a multilevel construct or a bivariate relationship indicates a wholes, parts, or equivocal situation when aggregated to one particular level. The purpose of MLA is to detect what happens to a multilevel construct or a bivariate relationship when aggregated to two (or more) nested levels – whether a situation (i.e., wholes, parts) applies to both levels (cross-level), only the lower level (level-specific), or only the higher level (emergent). For example, if a WABA analysis shows homogeneous individuals within classes (wholes), it is still possible for classes within schools to be wholes, parts, or equivocal. However, if the analysis shows heterogeneous individuals within classes (parts), then between-class variation is assumed to be error, and no higher level analyses are possible. The case of independent individuals within groups (equivocal) is interesting, as it is possible for the next higher level to show emergent properties of wholes or parts (Dansereau and Yammarino, 2000).

The purpose of MVA is to apply WABA techniques to a network of variables to describe the level at which each variable and bivariate relationship should be situated. Finally, MRA is used to identify any moderating variables that may affect inferences. These last two purposes relate most to the fifth step in the validation framework of Chen et al. (2004a), where the functional characteristics of the construct are tested across levels.

1.2.3 Comparison of the Two Frameworks

While Chen et al.'s (2004a) five-step validation framework is seen as a useful general approach to multilevel construct validation, its applicability to all situations has been questioned (Dansereau & Yammarino, 2004; Kim, 2004). They argue that constructs based on within-group variability cannot be modeled using the approach of Chen et al. (2004a) because of its traditional ANOVA perspective. In this perspective, groups either exist or they do not, and the definition of group membership is based on the assumption that all members are equal (i.e., equally treated). Thus, a group is homogeneous by definition, and the group mean applies equally to all members. Individual variation is viewed as unrelated to group membership, and treated as residual. This approach does not allow for the possibility of a heterogeneous group level of theory.

According to Dansereau and Yammarino (2004), the more modern statistical approach is a "sampling frame" perspective, which allows all three possible levels of theory. From this perspective, averaging observed scores makes sense when a group is composed of homogeneous individuals, while averaging deviation scores makes sense when a group is composed of heterogeneous individuals. When individuals are independent of group membership, averaging scores does not make sense. To be inclusive of potential within-group effects, Kim (2004) recommends the WABA analytic approach of Dansereau, Alutto, and Yammarino (1984) to multilevel construct validation, which focuses on the patterns of variability across levels.

In the context of multilevel construct validation, these statistical assumptions (i.e., ANOVA vs. sampling frame) will determine the applicability of different validation tests. For example, the reliability statistics suggested for steps 3 and 4 in the Chen et al. validation framework are valid estimators when the level of theory is homogeneous groups or independent individuals, but are invalid for heterogeneous group members (Kim, 2004). Likewise, attempts that have been made to apply statistical correction factors (Bliese & Halverson, 1998) to other

group-level estimators such as η^2 (used in WABA analyses) are flawed because they assume that within-group variation is error (Dansereau, Cho, & Yammarino, 2006).

1.3 School Readiness as a Population-Based Multilevel Construct

In this dissertation, these two multilevel construct validation frameworks will be applied to the construct of school readiness, in the specific form of an assessment tool called the EDI. The EDI will be described in detail in the next section. What is important to note now is that the EDI has been designed to reflect a particular view of school readiness as a population-based (vs. individual differences-based) multilevel construct. School readiness has not always been conceptualized (and therefore measured) in this way.

Traditional school readiness tests have a maturational/nativist theoretical perspective dating back to Gesell (1925). In this individual-differences view, school readiness implies that children have achieved the developmental stages necessary to enter existing school programs successfully. Rimm-Kaufmann and Pianta (2000) refer to this view as a "child effects model." Carlton and Winsler (1999) note that tests based on a maturational theory focus either on developmental milestones and/or on academic knowledge.

More contemporary theoretical approaches, such as transactional (Ford & Lerner, 1992), interactionist (Meisels, 1999), and ecological (Bronfenbrenner & Morris, 1998) emphasize the dynamic interactions between children and their various contexts (e.g., family, school, peers, neighbourhood). In these inherently multilevel perspectives, school readiness implies that both children and teachers/schools be ready for each other. The primary purpose of school readiness tests becomes one of ongoing assessment of appropriate programming during school, rather than assessing placement in advance of entering school (Carlton & Winsler, 1999).

Using school readiness tests to make inferences about individual children had its heyday in the U.S. in the mid- to late-1980s, falling into disrepute since then with educators, early childhood education (ECE) practitioners, and measurement specialists. There were several problems with readiness tests of that time. First, different jurisdictions developed and used a variety of tests that were usually administered in locally idiosyncratic ways, without any official definition of school readiness (Gnezda & Bolig, 1988; Mashburn & Henry, 2004; Saluja, Scott-Little, & Clifford, 2000). Second, certain widely used tests (e.g., Gesell Readiness Test, Brigance, DIALR, DABERON) were shown to have poor discriminant and/or predictive validity. For example, Ellwein, Walsh, Eads, and Miller (1991) found that a majority of children placed in extra-year programs on the basis of Gesell test results were misidentified.

Another consequence of this individual-differences perspective was the use of developmental screening tests and school readiness tests interchangeably for either special education referral or instructional planning (Gnezda & Bolig, 1988), when the former are designed specifically for referrals and the latter for instructional planning. Worst of all, delay of school entry or retention in kindergarten was sometimes justified on the basis of school readiness results, particularly for literacy and numeracy (Shepard, 1997). Voluntary delay of school entry, known as *redshirting*, was observed in the U.S., as parents (predominantly those with higher socioeconomic status [SES]) try to maximize the school success of their children. In these ways, the original intention of school readiness testing, which was to create more homogeneity in Grade 1, actually achieved greater heterogeneity in terms of age, SES, and pre-kindergarten education (Shepard, 1997).

Despite the poor showing of readiness tests in the 1980s, increasing investment in early childhood development and early intervention programs created renewed interest in the assessment of children entering school. In the past 20 years, readiness assessment has been

redefined, based on the 1989 U.S. National Educational Goals (National Education Goals Panel, 1991), which included the goal of universal school readiness in the U.S. by the year 2000. Three necessary characteristics were recommended for all school readiness assessments: coverage of at least four developmental areas (physical, social, emotional, and language/cognitive), appropriateness for children of that age, and achievement of high psychometric standards of reliability and validity (Mashburn & Henry, 2004).

Current directions in school readiness focus on its holistic nature as it relates to, and is influenced by, a diverse set of time-varying contextual influences at variety of ecological levels (Love, Aber, & Brooks-Gunn, 1994; Rimm-Kaufmann and Pianta, 2000). This emphasizes the importance and interrelatedness of peers, school, home, and neighbourhood as they collectively exert both direct and indirect effects on children over time. Such multilevel measures of school readiness that incorporate context in this way are important for good practice, policy, and research. As will be shown next, the EDI embraces this current direction in the conceptualization of school readiness, and extends it by purposefully restricting interpretations to the population level.

1.4 The EDI – Description, View of School Readiness, Level of Theory, and Summary of Validation Evidence

1.4.1 Description of the EDI

The EDI was developed starting in 1997 by Drs. Dan Offord and Magdalena Janus at the Canadian Centre for Studies of Children at Risk (now renamed the Offord Centre for Child Studies) at McMaster University. The development, testing and initial validation of the EDI was undertaken in partnership with the Founders' Network and the North York Early Years Action Group. Most of the 103 EDI core items were based on items in the National Longitudinal Survey

of Children and Youth (NLSCY), with additional items based on Doherty's (1997) review of the components of school readiness. These items together are designed to measure five areas or domains of child development: physical health and well-being (13 items), social competence (26 items), emotional maturity (30 items), language and cognitive development (26 items), and communication skills and general knowledge (8 items). The content of each item (i.e., what is being measured) can be seen in Chapter 2 as part of the descriptive summaries.

The EDI also includes additional items relating to children's special skills or talents (numeracy, literacy, arts, music, athletics/dance, problem-solving) and any special problems (physical, visual, hearing, speech, learning, emotional, behavioural, home environment). Finally, there are items about whether the child had prior experience in each of the following: an early intervention program, regular nonparental child care (by type of care), language or religion classes, an organized preschool/nursery school, and junior Kindergarten.

A number of child demographic characteristics are also recorded for each student by the kindergarten teacher. The child demographics collected are: class assignment (junior vs. senior kindergarten), date of birth, gender, home postal code, exceptional/special needs status, ESL status, French Immersion status, Other Immersion status, Aboriginal status, first language(s), adequacy of communication in first language, class membership status, and whether the child had been previously retained in kindergarten. On each child's form, teachers are also asked to provide demographic information about themselves: gender, age category, teaching experience (overall, at the school, at the kindergarten level, with that particular class), and educational attainment. Finally, two items about classroom characteristics are included: the number of children in the child's kindergarten class, and the class type (presence/absence of junior kindergarten, senior kindergarten, Grade 1 students in the class).

The EDI is administered by kindergarten teachers in February of the school year, after completing a training session on scoring the EDI items. Teachers complete the EDI instrument for each of the students in their kindergarten class. As of October 2007, EDI data had been collected for approximately 520,000 children across from all 10 provinces in Canada, including about 90,000 children in British Columbia (Janus et al., 2007). The EDI has also been used with minimal changes in a number of other countries, including Australia, Chile, Kosovo, Moldova, and Mexico.

The current version of the EDI is different from the pre-2005/06 versions in one important respect. The number of response categories (excluding the "don't know" response option) for 18 of the items was reduced from five to three, based on the recommendations by Andrich and Styles (2004) resulting from their Rasch Item Response Theory (IRT) analysis of the EDI. In addition, one binary item (measuring whether students can adequately communicate in their first language) was shifted from the Language and Cognitive Skills section to the section on child demographics.

1.4.2 EDI View of School Readiness

Of Meisel's (1999) four views of school readiness (idealist/nativist, empiricist/environmental, social constructivist, interactionist), Janus and Offord (2000, 2007) have chosen to ground the EDI in the social constructivist view. This perspective emphasizes community-level meanings and values, and requires broad-based measurement strategies that take a variety of child, family, and community contexts into account. This perspective also highlights the importance of community engagement with these measures for the purposes of developing, implementing, monitoring, and evaluating community-level interventions. EDI results are intended for interpretation only at a "community" level (e.g., school, district,

neighbourhood, community); child-level scores are never used or interpreted for any decision-making purposes for individual children.

The EDI is intended to measure children's readiness for Grade 1 entry, based on the five developmental domains described in section 1.4.1, rather than readiness at kindergarten entry. Janus and Offord (2007) provided two justifications for this choice. First, kindergarten is the time when children make the transition developmentally around the skills necessary for formal learning in school, and so these skills should be measured sometime during the kindergarten year. Second, measures of school readiness taken before kindergarten tend to be less predictive of future success in school.

1.4.3 EDI Level of Theory

The EDI is consistent with the notion of a multilevel construct. The creators of the EDI have consistently stated that it was designed exclusively for inferences at some level of aggregation of children, such as schools, school districts, or neighbourhoods. For example, Janus and Offord (2000, p. 74) wrote that "the EDI provides results at the population level…they are not interpretable for individuals." Their intent was to preclude anyone using EDI results to make programming, placement, entry, or retention decisions for individual children entering school, due to the historically substantial error rate. However, this also opens up the potential for committing atomistic fallacies.

The multilevel nature of the EDI is also inherent in the way that data are collected, using ratings by kindergarten teachers for each of the children in the class. One would expect a certain amount of between-classroom variation in EDI scores to be associated with teachers (as well as other aspects of the organizational context), as expressed by an intraclass correlation [ICC(1)] significantly greater than zero (Rowan, Raudenbush, & Sang, 1991). Indeed, Janus and Offord

(2007) have found intraclass correlations ranging from .21 to .31 for the five developmental scales measured by the EDI.

The EDI, with its population-based approach to school readiness, reflects recent theoretical trends in developmental science towards a "cell to society" or "neuron to neighbourhood" perspective (National Research Council Institute of Medicine, 2000) that incorporates research from diverse disciplines including neuro/psychobiology, health, education, psychology, epidemiology, and policy studies (Keating & Hertzman, 1999). The Human Early Learning Partnership (HELP) at the University of British Columbia, Canada is the home base for the EDI in the province. These school readiness scores provide an analytic base for early learning and development researchers representing many disciplines from six universities in British Columbia to create an environment of cross-discipline collaboration.

Do we know the level of theory for the EDI, at least in the context of kindergarten students within classrooms? The first option, homogeneous members, is certainly not the correct level of theory for the EDI. Developmental trajectories differ across children, and so there is no expectation of agreement or even similarity between children in a particular classroom. A much stronger case could be made that individual scores should be largely independent of classroom membership. After all, the intent of the EDI is to measure children's readiness for Grade 1, as "kindergarten provides the transition between the play-based preschool and home environment to the academically based environment of grade school" (Janus & Offord, 2007, p. 5). This level of theory, it seems, speaks to the appropriateness of inferences made at the individual child level. However, the EDI developers have consistently stated that it was designed exclusively for inferences at some level of aggregation of children, not for inferences at the individual child level.

An EDI level of theory based on interdependent scores is also a possibility. While all kindergarten teachers are rating students that they have known for only a few months, some have relatively small class sizes, and some have quite large class sizes. If this interdependent level of theory for the EDI is true, one likely implication is that the tendency to rate students relatively should be positively related to class size, at least above some threshold. For Cycle 1 EDI data for British Columbia, preliminary analyses show that this is indeed the case – where higher relativity in scoring is defined as an increasing ratio of within-class variation to between-class variation. For example, the proportion of total variation that is within-class for items in the social competence scale increased from 60% for classes with less than eight students to 82% for classes with more than 26 students. For the language and cognitive development scale, the corresponding increase was from 57% to 78%. The same general pattern is evident across all five scales.

1.4.4 EDI Validation Evidence

Until 2007, EDI validation studies (e.g., Duku & Janus, 2004; Janus, 2001; Janus, 2002; Janus, Offord, & Walsh, 2001; Janus, Walsh, & Duku, 2005; Janus, Willms, & Offord, 2000), were conducted entirely at the individual child level of analysis, without any multilevel considerations. In contrast, studies that have examined functional aspects of the EDI (e.g., how EDI scores associate with neighbourhood characteristics) have used both aggregated EDI scores in regression analyses (e.g., Kershaw & Forer, 2006) and individual EDI scores in HLM analyses (e.g., Carpiano, Lloyd, & Hertzman, in press; LaPointe, 2006).

The following sections summarize, and occasionally critique, the results of the reliability and validity studies conducted for the EDI from 2000 to 2007. With the exception of Guhn, Gadermann, and Zumbo's (2007) differential item functioning analysis (see section 1.4.4.9), and

Janus and Offord's (2007) multilevel covariance structure analysis (see section 1.4.4.10), these reliability and validity studies have relied on the weaker correlation-based validity methods.

1.4.4.1 Internal Consistency Reliability

Janus et al. (2000) conducted the first reliability and validity analyses of the five EDI scales. Using data from more than 16,000 kindergarten children in Ontario, there was good internal consistency for each of the five scales, with coefficient alphas ranging from .84 for the physical health and well-being scale to .96 for the social competence scale. A replication using a normative sample of more than 125,000 children (Janus, Walsh, & Duku, 2005) confirmed the high internal consistency of the scales.

At the subscale level, where the number of items is between four and nine, internal consistency was also high for all but one subscale. Using the same normative sample, Janus, Walsh, and Duku (2005) found coefficient alpha values ranging from .75 for the basic literacy skills subscale to .94 for the prosocial and helping behaviour subscale, The one subscale with low internal consistency was the physical independence subscale, with a coefficient alpha of .26.

1.4.4.2 Interrater Reliability

Interrater reliability was measured (Janus et al., 2000; Janus & Offord, 2007) by comparing teachers' and early childhood educators' EDI scores for 53 children. Moderate to high correlations (ranging from .53 to .80 for the five scales) suggested to the authors that "the concepts captured by the EDI are clear and are easily assessed by trained educators" (Janus & Offord, 2007, p. 15). Janus (2001) also tested the interrater reliability of the EDI using a different sample of 51 children, with the EDI completed for each child by the kindergarten teacher, the child care teacher, and a parent. The resulting correlations between kindergarten teacher and

child-care teacher were essentially identical in range to those found by Janus et al. (2000). Interrater reliabilities involving parents were lower, though still statistically significant – kindergarten teacher and parent ratings had correlations ranging from .35 to .66, while child care teacher and parent ratings had correlations ranging from .26 to .52.

1.4.4.3 Intrarater Reliability

Duku and Janus (2004), for a very small sample of teachers (*n*=5 in Saskatchewan and *n*=9 in Hamilton), measured the within-teacher reliability over approximately a three-week period using an intraclass correlation coefficient. Mean within-teacher reliabilities ranged from .70 to .95, but the intrarater correlation was as low as .18, for the emotional maturity scale.

1.4.4.4 Test/Retest Reliability

Test-retest reliability, not surprisingly, has been assessed with small samples of children. Over a four-week period for a sample of 112 students, Janus et al. (2000) found test-retest correlations to be high (.82 to .94) and statistically significant. Duku and Janus (2004) also looked at test-retest reliability (with a two- to five-week interval) for EDI samples of similar size, in Saskatchewan and Ontario. Test-retest correlations were greater than .76 for all subscales and for each community.

1.4.4.5 Validity – Factor structure

Until recently, the dimensionality of the EDI has only been explored by researchers at the Offord Centre for Child Studies, where the EDI was originally developed. Their original factor analytic approach was to use principal components analysis (PCA) with varimax rotation, using data at the individual child level. The first such PCA (Janus et al., 2000) using a sample of over

16,000 children, resulted in an 11 factor solution "which could be clearly aggregated into the five [theoretical developmental] domains" (p. 6). However, 17 factors had eigenvalues greater than 1.0. Factor loadings of 0.3 or higher was the stated criterion for retaining an item in a factor. Five items with loadings less than 0.3 were retained, however, on the basis of perceived importance by teachers. The authors also stated that several items were removed as a result of low communalities or factor loadings, but it is not clear how many items, or which ones. This removal also seems to have been temporary, as the items in a later (2005) version of the EDI were the same as those in 2000 (with response category changes only).

A replication of the PCA using the large-scale normative sample (Janus, Walsh, & Duku, 2005) was deemed to have confirmed the existence of the same five distinct developmental scales, as well as subscales within four of the five scales, for a total of 16 dimensions at the subscale level. The authors claimed that these same 16 subscales had previously been identified and confirmed earlier, with 1999/2000 and 2000/2001 data respectively, but there is no documentation to confirm this. Consistent with Janus et al. (2000), each item was assigned to a particular subscale on the basis of factor loadings, but with conceptual considerations occasionally taking precedence. For example, three items in the language and cognitive development scale were assigned to a subscale called "basic literacy skills" despite having small factor loadings on that dimension, ranging from 0.04 to 0.15.

Using the same sample of approximately 16,000 children from the Janus et al. (2000) unpublished study, Janus and Offord (2007) conducted a reanalysis of the EDI factor structure, this time using principal axis factoring with promax rotation. Although they described this reanalysis as "confirmatory," it was in fact an exploratory factor analysis, as they used the criterion of eigenvalues greater than 1.0 to determine the number of retained factors. Fourteen factors were retained and then assigned to the five theoretical domains. As in their previous

factor analyses, some items (seven in this case) were assigned to factors despite having factor loadings under 0.3, on the basis of perceived importance by teachers.

Janus and Offord (2007) extended this examination to consider the multilevel factor structure of the EDI. These results are summarized and discussed in section 1.4.4.10.

1.4.4.6 Validity – Concurrent

According to Janus and Offord (2007), there is evidence for EDI concurrent validity based on modest correlations between PPVT and EDI language and cognitive development scores (r = .31) and PPVT and EDI communication and general knowledge scores (r = .47) for a sample of 82 children. In a smaller study of 51 children, Janus (2001) found slightly higher concurrent validity between the PPVT and EDI language and cognitive development scores (r = .44) and between the PPVT and EDI communication and general knowledge scores (r = .49). In contrast, Janus, Offord, and Walsh (2001), using a sample of 480 children, found a more modest relationship between the PPVT and EDI language and cognitive scores (r = .26), though a similar correlation between the PPVT and EDI communication and general knowledge scores (r = .57). Concurrent validity, as measured by the correlation between PPVT and EDI subscale scores, was lower when the EDI was rated by child care teachers (Janus, 2001).

Brinkman, Silburn, Lawrence, Goldfield, Sayers, and Oberklaid (2007) investigated the concurrent and construct validity of the Australian EDI (AEDI) for a sample of 642 children by comparing AEDI scores with scores from the Longitudinal Survey of Australian Children (LSAC). The authors found moderate to high correlations between AEDI subscales and similar teacher-rated LSAC measures, which included the PPVT, Who Am I, Parents Evaluation of Developmental Status, and the Strengths and Difficulties Questionnaire.

EDI concurrent validity has also been assessed for a sample of 151 Jamaican children, comparing associations between two of the EDI scales (language and cognitive development, and communication and general knowledge) and two measures of language and cognitive skills, the McCarthy Scales of Children's Abilities and the PPVT. Correlations were moderate (.31 to .45) and statistically significant (Janus et al, 2007).

1.4.4.7 Validity – Discriminant

According to Janus et al. (2000), PPVT scores and the three EDI scales not related to language or communication (i.e., physical health and well-being, social competence, and emotional maturity) were not significantly correlated (r's from .01 to .13), which was provided as evidence of discriminant validity.

1.4.4.8 Validity – Predictive

In British Columbia, predictive validity has been assessed by linking school readiness (EDI) scores in kindergarten to province-wide school competency scores in Grade 4 (called the Foundation Skills Assessment). The Personal Education Number (PEN) for each student provides the necessary link. The number of EDI scales for which children were vulnerable in kindergarten predicts, in a dose-response fashion, lack of school competence in Grade 4 for numeracy, reading, and writing (Janus et al., 2007). In Quebec, EDI scores were found to predict 36% of the variance in Grade 1 school achievement scores, and 72% of the total variance explained in combination with a battery of other cognitive measures (Forget-Dubois et al., 2007). In Ontario, EDI subscale scores showed low to moderate, but statistically significant, correlations with three instruments (Developmental Test of Visual-Motor Integration, Strengths and Difficulties Questionnaire, Detroit Test of Learning Aptitude) measured three years later.

Janus, Harren, and Duku (2004) found small but statistically significant associations between EDI scores and Grade 3 reading and writing scores, at both the school and neighbourhood levels.

1.4.4.9 Sub-Group and Differential Item Functioning Analyses

Janus (2002) conducted a validation study for Aboriginal children, using data from the original Understanding the Early Years (UEY) sites in Saskatchewan and Manitoba. Development-related measures captured for these UEY children were the EDI, two direct assessments (PPVT and Who Am I), and child behaviour subscales (hyperactive-inattention, emotional disorder/anxiety, aggression) from the parent questionnaire of the NLSCY. The direct assessments and behaviour subscales were used to test the concurrent validity of the EDI, using a subsample of 45 Aboriginal and 45 non-Aboriginal children matched on household income, parent education, parent age, and parent gender. These matched children represented the top of the Aboriginal distribution and the bottom of the non-Aboriginal distribution. For both subsamples, the highest correlations with the PPVT were for the language and cognitive development, and communication and general knowledge subscales (Pearson r from .29 to .40). All of the EDI subscales had a low correlation with Who Am I scores (highest r = .18). These low correlations were different than found by Janus, Offord, and Walsh (2001), where the Who Am I scores were correlated at .35 or higher with three EDI subscales – social competence, emotional maturity, and language and cognitive development. The highest correlations between an EDI subscale and child behaviours were for both EDI social competence and emotional maturity with the NLSCY hyperactivity-inattention score (both over .3, for both Aboriginal and non-Aboriginal children). Despite the matching, Aboriginal children scored significantly worse than non-Aboriginal children on PPVT and on EDI subscales for physical health and well-being,

emotional maturity, and communication and general knowledge; this is an apparent reflection of the disadvantage of Aboriginal children in their readiness for school at Grade 1.

Guhn, Gadermann, & Zumbo (2007) used differential item functioning (DIF) analysis to examine potential measurement bias due to child gender, English as a Second Language (ESL) status, and Aboriginal status. There were no differences by Aboriginal status. There was a significant gender difference for only one item, with boys more likely rated as physically aggressive compared to girls. ESL students were scored lower on items in the language and cognitive development, and communication and general knowledge scales.

1.4.4.10 Multilevel Construct Validation of the EDI

Janus and Offord's (2007) article represents the first attempt to conduct a multilevel construct validation of the EDI. They considered two levels (individuals and classes), and looked at three aspects of the multilevel construct: the amount of variation attributable at the class level, the reliability of the class means, and the factor structure. These were tested by calculating the intraclass correlation coefficient [ICC(1)], calculating the generalizability coefficient [ICC(2)], and conducting a multilevel confirmatory factor analysis (Muthen, 1994), respectively. On the basis of their results, the authors concluded that classroom level variation was minor, that class means were reliable, and that the factor structure was identical across the two levels.

These conclusions are questionable on at least two fronts. First, depending on the developmental domain involved, the intraclass correlation ranged from .21 to .31 in their sample, much too large for between-group variance to be considered as error. In more mainstream interpretation of ICC(1) results, values as low as .01 can be sufficient to allow significant effects to emerge at the aggregate level (Bliese, 2000), as long as group means are reliable, as they were in their study.

Second, Janus and Offord's (2007) assessment of the multilevel (individual and classroom) factor structure of the EDI is debatable. They purported to conduct a multilevel covariance structure analysis using Muthen's (1994) procedure. However, it is unclear how the reported analyses correspond to the steps in that approach. Also, given that they report CFI, RMSEA, and SRMR fit statistics, their analysis must have assumed that EDI item scores are continuous, rather than categorical. Although the multilevel confirmatory factor analysis showed similar CFI values at each of the two levels, overall model fit to the data was poor in both cases, undermining any claims about the factor structure at either level, much less similarity across levels.

1.5 Teacher Effects

There are two basic approaches to measuring aspects of children's early learning and development: direct assessment (especially using standardized tests), and teacher ratings. The former approach (e.g., PPVT-R, Who Am I) has the decided validity advantage of known psychometric properties, particularly relevant given the inherent limitations in young children's response capacity. However, such tests are expensive to administer, are not practical as universal assessment tools, and do not have a ready connection to children's formal learning environments such as to school curricula.

In contrast, teacher ratings (like those used for the EDI) can be gathered efficiently for all children in a classroom, and have the contextual advantages of relevance and knowledge about students. However, potential teacher bias is a particular concern precisely because of this contextual relevance. Teacher ratings on readiness tests can and do result in real-life developmental trajectory-altering consequences for individual children, such as referrals to special services, remedial placements, and grade retention. This has been especially salient in the

U.S., where kindergarten teachers are often expected to be gatekeepers in response to increasing curricular and accountability expectations (Shepard, 1997).

Even when teacher-rated school readiness tests are not interpreted at the individual child level (as is mandated for the EDI), there are still consequences for potential harm at other ecological levels. For example, if teachers in an inner-city school decide to take the low socioeconomic status of their students into account when rating their students' early literacy, that school may not be targeted for inclusion in a new district-wide bookmobile program, based on the unexpectedly high literacy scores.

Messick (1989) defines teacher bias as systematic construct-irrelevant differences in the ways that teachers view the construct that they are rating. There are several potential components of teacher bias, such as student characteristics (e.g., gender, race, family SES), teacher characteristics (e.g., ability, severity, consistency over time), assessment characteristics (e.g., task difficulty, rating scale structure), and interaction effects between each of these types of components (Englehard, 2002).

The most widely studied teacher effects in educational research are those related to student characteristics. Perhaps the most well-known and reliable effects are those found in the Pygmalion in the Classroom studies (Rosenthal & Jacobson, 1992), which established that teacher expectations can affect students' intellectual performance. Students' racial and cultural characteristics have been shown to bias teacher ratings of student achievement and behavioural problems (e.g., Epstein et al., 2005; Neal, McCray, Welk-Johnson, & Bridgest, 2003; Serwatka, Dove, & Hodge, 1986). Teacher biases relating to student gender and family characteristics like mother's marital status and education have also been shown with regard to the assessment of early reading problems (e.g., Beswick, Sloat, & Willms, 2003; Shaywitz, 2003). Finally, in direct comparisons of school readiness ratings by preschool teachers and kindergarten teachers,

the latter were found to be more influenced by the race of the child and the family's economic circumstances (Mashburn & Henry, 2004).

There are comparatively fewer teacher effects studies that focus on teacher and assessment characteristics. Saal, Downey, and Lahey (1980) found that the most frequent rater errors are severity/leniency, halo effects, response sets, and restriction of range. Andrich and Styles (2004), using a Rasch model analysis of Australian kindergarten teachers' EDI ratings, showed that teachers tended to restrict their range of answers for 18 of the items. Their recommendation to reduce the number of response categories for these items was adopted for both the Australian and Canadian EDI starting with the 2005 version of the instrument. Mashburn, Hamre, Downer, and Pianta (2006) found, for pre-kindergarten children, that teacher ratings of social competence and problem behaviours (similar to the emotional maturity scale for the EDI) were associated with teacher and classroom characteristics such as teacher experience and class size.

The predominant attention on student characteristics in teacher effects studies is consistent with the traditional individual-differences paradigm in educational research, in that both the outcome and explanatory variables are measured at the individual child level. On the other hand, when examining the role of teacher characteristics on their ratings of individual children, a multilevel model is necessary. Mashburn et al. (2006), for example, used HLM in their analysis of teacher and classroom characteristics on their ratings. Given that the purpose of this dissertation revolves around multilevel construct validation, the examination of teacher effects will focus on the influence of teacher and classroom characteristics, rather than child characteristics.

1.6 Three Research Questions

As set out in the introduction, the purposes of this dissertation are to: 1) articulate and develop the Chen et al. and WABA multilevel construct validation frameworks, and 2) apply each in the context of the EDI. The first purpose has partially been met in the literature review, and will be further explored in Chapters 3 and 4 where the frameworks will be discussed specifically in the context of the EDI. This sets the stage for the second purpose, where the frameworks' analytic techniques are adapted and applied to EDI data from British Columbia. Three research questions will be answered in the process of meeting this second purpose. The first question focuses on the structural aspects of the construct validity of the EDI, while the second and third questions address functional aspects relating to teacher and classroom influences on EDI scoring.

Research Question 1

Based on the two validation frameworks, what is the evidence regarding the construct validity of EDI scale scores at different levels of aggregation, using both actual and vulnerability scores?

Research Question 2

To what extent do teacher and classroom characteristics affect the multilevel factor structure for each EDI scale?

Research Question 3

To what extent do teacher and classroom characteristics moderate the relative amounts of within-class and between-class variance for each EDI scale?

Before these questions are addressed, information about the study sample and descriptive summaries of EDI scores at various levels of aggregation are provided in the next chapter.

Chapter 2

Study Sample and Descriptive Summaries of EDI Scores and Teacher/Classroom Characteristics

2.1 Study Sample

The data used in this study come from HELP's second round (i.e., Cycle 2) of provincial EDI data collection, which began in 2004 and was completed in the spring of 2007. A complete Cycle 2 provincial dataset was not yet available when the analyses were conducted for this dissertation, but the available data represent 54 of the 59 school districts in the province. Overall, there are data for 36,407 children, representing 466 of the 478 HELP-defined neighbourhoods, and 2,497 classrooms.

It should be noted that no sampling weights have ever been used for the EDI, in British Columbia or elsewhere, and so none of the results below have been weighted. Essentially, EDI data constitutes a very large convenience sample, with almost complete coverage (Janus et al., 2007). The mean coverage over Cycles 1 and 2 was 98% and the median coverage was 93%.

One important aspect of this dissertation is an examination of the effects of teacher and classroom characteristics on inferences. For children with missing teacher characteristics, these variables were missing entirely in more than 90% of the cases. Therefore, there was no reasonable way to impute teacher characteristics, and so all of the analyses were based only on the 26,005 individual child records in the dataset where all of the teacher and classroom information on the EDI form was complete. This also excludes the 116 classrooms in the sample where there were two kindergarten teachers.

However, 26,005 is not the actual sample size for each scale used in the analyses, due to different amounts of missing item data for each scale. For each scale separately, cases were deleted if they had any items with missing data. The level of missing data was highest for the

emotional maturity scale, where the effective N was reduced to 18,185. The number of individual children for each scale is shown in Tables 1 to 5 below.

2.2 Descriptive Summaries of EDI Item Scores, All Scales, Individual Level

Table 1 shows, for each item in the physical health and well-being scale, the proportion of children in each response category, based on their teacher ratings. This scale is unique in that it contains a mixture of binary and three-category ordinal items. For the other four scales (see Tables 2 to 5), all items are either binary (language and cognitive scale) or three-category ordinal (social competence, emotional maturity, communication and general knowledge scales). For the EDI, binary items are scored as either 0 or 10, while the three-category items are scored as 0, 5, or 10.

Table 1 Physical Health and Well-Being Scale Items, Proportion for Each Category, Individual Level (N=21,504)

Item	***Yes***	***No***	
A2. Over or underdressed for school-related activities	6.3	93.7	
A3. Too tired/sick to do school work	8.9	91.1	
A4. Late	25.8	74.8	
A5. Hungry	4.0	96.0	
A6. Independent in washroom activities most of the time	98.6	1.4	
A7. Shows an established hand preference	97.3	2.7	
A8. Is well-coordinated	92.7	7.3	
	Poor/ v. poor	***Average***	***Good/ v. good***
A9. Proficiency at holding pen, crayons, or brush	10.0	37.6	52.4
A10. Ability to manipulate objects	4.2	39.8	56.0
A11. Ability to climb stairs	2.6	39.9	57.4
A12. Level of energy throughout the school day	3.6	40.5	55.9
A13. Overall physical development	2.8	43.0	54.2
	Often	***Sometimes***	***Never***
C58. Sucks a thumb/finger	1.3	3.1	95.5

Table 2 Social Competence Scale Items, Proportion for Each Category, Individual Level (N=24,163)

Item	***Poor/ v. poor***	***Average***	***Good/ v. good***
C1. Overall social/emotional development	9.8	44.2	46.0
C2. Ability to get along with peers	7.7	42.5	49.8
	Never	***Sometimes***	***Often***
C3. Plays/works cooperatively with other children	3.1	29.0	67.9
C4. Is able to play with various children	3.9	30.3	65.8
C5. Follows rules and instructions	2.4	28.3	69.3
C6. Respects the property of others	1.9	18.3	79.8
C7. Demonstrates self-control	4.0	29.7	66.3
C8. Shows self-confidence	4.9	36.9	58.2
C9. Demonstrates respect for adults	1.3	17.0	81.7
C10. Demonstrates respect for other children	1.9	25.9	72.1
C11. Accepts responsibility for actions	4.7	26.3	69.0
C12. Listens attentively	5.4	37.9	56.6
C13. Follows direction	3.1	31.1	65.8
C14. Completes work on time	5.2	26.9	67.9
C15. Works independently	5.6	26.1	68.3
C16. Takes care of school materials	1.9	18.8	79.3
C17. Works neatly and carefully	6.5	32.5	61.0
C18. Is curious about the world	1.5	19.0	79.5
C19. Is eager to play with a new toy	0.8	14.5	84.7
C20. Is eager to play a new game	1.5	16.9	81.6
C21. Is eager to play with/read a new book	3.4	22.3	74.3
C22. Is able to solve day-to-day problems by self	8.2	46.6	45.3
C23. Is able to follow one-step instructions	1.5	18.6	79.9
C24. Able to follow class routines without reminders	3.9	29.4	66.7
C25. Is able to adjust to changes in routines	2.9	23.0	74.1
C27. Shows tolerance to someone making a mistake	3.6	29.5	66.9

Table 3 Emotional Maturity Scale Items, Proportion for Each Category, Individual Level (N=18,185)

Item	***Never***	***Sometimes***	***Often***
C28. Willing to try to help someone who is hurt	10.4	39.1	50.5
C29. Volunteers to help clear mess made by other	20.2	42.6	37.2
C30. Will try to stop quarrel or dispute	30.5	45.4	24.2
C31. Offers to help others having task difficulty	23.0	42.7	34.3
C32. Comforts a child who is crying or upset	21.3	43.4	35.3
C33. Spontaneously picks up others' dropped objects	24.0	44.3	31.6
C34. Invites bystanders to join in a game	25.6	50.0	24.4
C35. Helps other children who are feeling sick	24.9	44.4	30.7
C36. Is upset when left by parent/guardian	81.3	14.7	4.0
C37. Gets into physical fights	88.4	10.0	1.6
C38. Bullies or is mean to others	84.2	13.7	2.1
C39. Kicks, bites, hits other children or adults	90.3	8.2	1.5
C40. Takes things that do not belong to him/her	90.3	8.4	1.3
C41. Laughs at other children's discomfort	84.9	13.8	1.2
C42. Can't sit still, is restless	69.5	23.0	7.5
C43. Is distractible, has trouble sticking to activities	71.8	20.6	7.6
C44. Fidgets	67.5	24.4	8.1
C45. Is disobedient	81.4	15.7	2.9
C46. Has temper tantrums	88.9	9.4	1.8
C47. Is impulsive, acts without thinking	68.7	25.2	6.0
C48. Has difficulty waiting turn in games or groups	70.4	23.8	5.9
C49. Can't settle on anything for any length of time	81.6	14.7	3.6
C50. Is inattentive	60.6	32.4	7.0
C51. Seems to be unhappy, sad or depressed	81.1	16.4	2.5
C52. Appears fearful or anxious	81.6	16.0	2.4
C53. Appears worried	77.1	20.4	2.5
C54. Cries a lot	89.2	9.1	1.8
C55. Is nervous, high-strung, or tense	85.3	12.3	2.4
C56. Is incapable of making decisions	80.8	16.6	2.6
C57. Is shy	64.2	28.2	7.6

Table 4 Language and Cognitive Development Scale Items, Proportion for Each Category, Individual Level (N=20,472)

Item	***No***	***Yes***
B8. Knows how to handle a book	0.9	99.1
B9. Is generally interested in books	4.4	95.6
B10. Is interested in reading	11.2	88.8
B11. Able to identify at least 10 letters of the alphabet	11.8	88.2
B12. Is able to attach sounds to letters	21.8	78.2
B13. Is showing awareness of rhyming words	17.8	82.2
B14. Is able to participate in group reading activities	9.5	90.5
B15. Is able to read simple words	32.5	67.5
B16. Is able to read complex words	80.5	19.5
B17. Is able to read simple sentences	58.6	41.4
B18. Is experimenting with writing tools	12.5	87.5
B19. Is aware of writing direction in English	8.1	91.9
B20. Is interested in writing voluntarily	39.8	60.2
B21. Is able to write his/her name in English	3.7	96.3
B22. Is able to write simple words	20.1	79.9
B23. Is able to write simple sentences	45.7	54.3
B24. Is able to remember things easily	16.6	83.4
B25. Is interested in mathematics	11.2	88.8
B26. Is interested in games involving numbers	11.0	89.0
B27. Is able to sort and classify objects	4.3	95.7
B28. Is able to use on-to-one correspondence	6.8	93.2
B29. Is able to count to 20	12.8	87.2
B30. Is able to recognize 1 – 10	13.5	86.5
B31. Is able to say which of two numbers is bigger	13.7	86.3
B32. Is able to recognize geometric shapes	6.3	93.7
B33. Understands simple time concepts	7.2	92.8

Table 5 Communication and General Knowledge Scale Items, Proportion for Each Category, Individual Level (N=24,774)

Item	***Poor/ v. poor***	***Average***	***Good/ v. good***
B1. Ability to use language effectively in English	12.3	36.2	51.6
B2. Ability to listen in English	6.1	39.1	54.8
B3. Ability to tell a story	15.1	40.5	44.4
B4. Ability to take part in imaginative play	6.1	41.3	52.6
B5. Ability to communicate own needs understandably	11.0	37.6	51.4
B6. Ability to understand something on the first try	9.9	36.7	53.4
B7. Ability to articulate clearly	14.0	37.0	49.0
	Never	***Sometimes***	***Often***
C26. Answers questions showing knowledge of world	5.2	21.4	73.5

2.3 Treatment of Missing Data

When there is more than a trivial amount of missingness, as is the case for the EDI data for this dissertation, the technique of deleting all cases with any missing data is typically not recommended (Schafer & Graham, 2002). Multiple imputation or adaptive maximum likelihood estimation (which is a sort of implicit multiple imputation) are unbiased and efficient missing data techniques when missingness is at random or completely at random (Allison, 2002). However, as shown below, missingness was not likely at random in the current circumstances, as is expected for a paper administration of the EDI.

The decision to restrict empirical analyses to children with complete EDI data was made in the context of construct validation. EDI data collection is designed to take place about halfway through the school year so that all teachers theoretically have had sufficient opportunity to make informed judgments for all of their students on all 103 EDI items. For the complete-data cases in the sample, informed judgments are assumed, and therefore the instrument is being used as designed.

When EDI data are missing, regardless of whether this is because of "don't know" or blank responses, a likely explanation is that the teacher did not feel confident about making an informed judgment. If this lack of confidence only applied to the items with missing data, then one would expect that mean scale scores would still be the same for children with and without missing data. However, as Table 6 shows, for all EDI scales except emotional maturity, mean scale scores were lower for children with one or more missing items, although the effect sizes are small at best (Cohen, 1988). These results, therefore, may reflect a slight negative bias in scoring for children about whom teachers generally do not feel as well informed.

Therefore, in conducting a construct validation for the EDI, cases were included only if teachers used the instrument as designed, with complete data. If a multiple imputation or

adaptive maximum likelihood technique was used, a higher probability of inferential error could be the result, because it would tend to add cases where scores may be negatively biased. Happily, the sample sizes with complete data are still very large, and so the deletion, does not affect the validation analyses to follow.

Table 6. Comparing Cases With/Without Missing Data, by EDI Scale

	Items Missing	***No Items Missing***	***t-test (df)***	***Effect Size (η^2)***
	Mean (SD)	***Mean (SD)***		
Physical	8.21 (1.53)	8.59 (1.40)	15.5* (25446)	.009
Social	6.95 (2.39)	8.23 (1.87)	23.7* (25841)	.021
Emotional	7.92 (1.43)	7.91 (1.61)	0.5 (25169)	.000
Language	7.86 (2.03)	8.14 (1.96)	9.1* (25339)	.003
Communication	5.32 (2.71)	7.23 (3.16)	18.8* (25515)	.014

Note: * $p < .001$.

An attrition bias analysis was conducted to examine the extent to which being excluded from the analysis because of missingness related to the following child characteristics: gender, ESL status, special needs status, and Aboriginal status. The 2x2 chi-squared values by themselves are relatively uninformative, given the very large sample sizes. However, effect sizes (based on the phi coefficient) were small across these child characteristics and the five scales (Cohen, 1988). All effect sizes were under .06, with the exception of the social competence scale for special needs children. In that case, there were missing values for 23% of children with special needs, compared to 6% of children without special needs, and the corresponding phi coefficient was .12.

2.4 Descriptive Summaries of EDI Scale Scores, Various Levels of Aggregation

Table 7 shows the descriptive summary statistics for each scale at the individual child level. As previously discussed, the two common methods for summarizing EDI scale scores have been employed in this dissertation. In the first method, each child receives an EDI scale score that is the mean of item scores in that domain for that child. The descriptive statistics at the individual level summarize the distribution of the EDI scale scores across children. In the second method, each child is assigned a binary vulnerability category (vulnerable or not) for each domain. Children are categorized as vulnerable for a scale if their score falls in the bottom decile of scale scores for children across British Columbia, based on the almost 44,000 children from Cycle 1 (1999 to 2004) of the EDI (Kershaw, Irwin, Trafford, & Hertzman, 2005). The descriptive statistics for EDI vulnerability scores at the individual level summarize the proportion of children in the sample who are identified as vulnerable. The proportion is not exactly 10% for each scale because a) the results are based on a different sample of children (i.e. Cycle 2), and b) missing data can affect the proportions.

Table 7 Descriptive Statistics for All EDI Scale Scores, Individual Level

	Number of Students	***Mean***	***Standard Deviation***
Physical – actual scores	21,504	8.59	1.40
Social – actual scores	24,163	8.23	1.87
Emotional – actual scores	18,185	7.91	1.61
Language – actual scores	20,472	8.14	1.96
Communication – actual scores	24,774	7.23	2.71
Physical – vulnerability scores	21,504	0.10	0.30
Social – vulnerability scores	24,163	0.12	0.32
Emotional – vulnerability scores	18,185	0.12	0.33
Language – vulnerability scores	20,472	0.11	0.31
Communication – vulnerability scores	24,774	0.12	0.32

Given that the focus of this dissertation is on multilevel validation methods, descriptive statistics are also provided at the two levels of aggregation that are explored in the forthcoming analyses and results. Table 8 provides statistical summaries for each scale at the classroom/teacher level, while Table 9 provides statistical summaries at the neighbourhood level. Not surprisingly, the relative values of the mean scores (e.g., physical health and well-being highest, and communication and general knowledge lowest) is preserved, regardless of the level of aggregation. Similarly, the standard deviations show the typical pattern of ever-smaller standard deviations as the level of aggregation (i.e., group size) increases.

Table 8 Descriptive Statistics for All EDI Scale Scores, Class Level

	Number of Classes	***Mean***	***Standard Deviation***
Physical – actual scores	1,555	8.54	0.84
Social – actual scores	1,688	8.18	0.99
Emotional – actual scores	1,508	7.86	1.02
Language – actual scores	1,628	8.11	1.20
Communication – actual scores	1,692	7.50	1.60
Physical – vulnerability scores	1,555	0.11	0.15
Social – vulnerability scores	1,688	0.12	0.15
Emotional – vulnerability scores	1,508	0.13	0.15
Language – vulnerability scores	1,628	0.11	0.16
Communication – vulnerability scores	1,692	0.12	0.16

Table 9 Descriptive Statistics for All EDI Scale Scores, Neighbourhood Level

	Number of Neighbourhoods	***Mean***	***Standard Deviation***
Physical – actual scores	410	8.57	0.46
Social – actual scores	412	8.27	0.62
Emotional – actual scores	407	7.96	0.57
Language – actual scores	407	8.17	0.73
Communication – actual scores	412	7.29	1.03
Physical – vulnerability scores	410	0.10	0.10
Social – vulnerability scores	412	0.11	0.10
Emotional – vulnerability scores	407	0.12	0.10
Language – vulnerability scores	407	0.11	0.10
Communication – vulnerability scores	412	0.10	0.09

2.5 Descriptive Summary of Teacher and Classroom Characteristics

On the EDI forms, information for several classroom and teacher characteristics was recorded, of which five were used in the two functional analyses that follow. These five variables were: the number of children in the classroom, teacher months of experience at the kindergarten level, teacher age, teacher educational attainment, and teacher gender. Two of these variables, the number of children in the classroom and the number of months of experience as a kindergarten teacher, were coded as continuous variables. Teacher age group was coded in 10-year intervals, with *20 to 29 years old* as the lowest category and *60 years old and above* as the highest category. An interval level of measurement was assumed for this variable. The educational attainment of the teachers was the most complex variable, as it was coded as an 11-category multiple-response variable. This was simplified to three mutually exclusive ordinal categories: *no Bachelor degree, Bachelor degree but no graduate training*, and *some graduate training or attainment of a graduate degree*.

The descriptive statistics at the classroom level for these five covariates are shown in Table 10. There were 15.3 students in the average class. A large majority of teachers were female. The average teacher was 40 years old, and had been teaching at the kindergarten level for 103 months (about 8.5 years). The most common highest educational attainment was a Bachelor degree, with 16% having additional graduate training or attaining a graduate degree.

These descriptive statistics are a summary for all 1,702 teachers/classrooms where there was one teacher and complete teacher/classroom information was recorded. The number of teachers/classrooms for each of the analyses for Research Questions 2 and 3 will be lower and scale-specific, due to variations in of the proportion of missing EDI scores for each scale. Table 8 above indicates the number of classrooms with available teacher/classroom information for each EDI scale.

Table 10. Descriptive Statistics for Classroom and Teacher Characteristics, Classroom Level (N= 1,702)

		Mean or Percentage	***Standard Deviation***
Number of students in the class		15.3	6.1
Teacher gender	Female	98%	--
Teacher gender	Male	2%	--
Teacher age (years)		40.0	10.5
Teacher kindergarten experience (months)		102.9	93.0
Teacher education	No Bachelors	8%	--
Teacher education	Bachelor	76%	--
Teacher education	Graduate training	16%	--

Chapter 3

Background, Methods, and Results for Research Question 1

Research Question 1: Based on the two validation frameworks, what is the evidence regarding the construct validity of EDI scale scores at different levels of aggregation, using both actual and vulnerability scores?

3.1 Background – Applying Frameworks to EDI

The first order of business in considering the validity of the EDI relative to the multilevel construct framework of Chen et al. (2004) and the WABA approach of Dansereau et al. (1984) is to delineate the appropriate types of validation evidence for various ways of aggregating individual EDI scores.

3.1.1 Chen et al. (2004a) Framework

The first step in the Chen framework deals with the theoretical issues of construct definition: the domain boundaries of the construct, its dimensionality, whether it is a measure or an index (or elements of both), and its essential nature (i.e., how best to summarize the scores). The resolution of these definitional issues is important, as it provides the basis for moving to the next steps in the framework. In this dissertation, current construct definitions are assumed in the interest of keeping the scope of the dissertation research manageable. However, definitional work for the EDI is still needed, as will be shown next.

The original EDI domain boundaries (as represented by the 103 items) have been firmly set since its original construction, despite some contrary evidence. The process by which the EDI was created and pretested was thorough, as described in Janus and Offord (2007). However,

some items continue to be retained despite low factor loadings (Janus, Walsh, & Duku, 2005) and/or Rasch model misfit (Andrich & Styles, 2004), with retention justified in terms of meeting teachers' domain expectations (Janus & Offord, 2007). Regarding dimensionality, the EDI was devised to cover five developmental domains: physical health and well-being, social competence, emotional maturity, language and cognitive development, and communication and general knowledge. Not surprisingly, these dimensions have been asserted in various principal components or principal axis factoring analyses, despite a few items having higher loadings on a different dimension than theorized (Janus & Offord, 2007).

There has been no discussion in the EDI literature about whether EDI scales are theorized to be measures or indices. However, the implicit assumption seems to be that they are measures, given that Janus, Walsh, and Duku (2005) assessed the reliability of each scale by calculating coefficient alpha coefficients. As Bollen and Lennox (1991) have shown, the need for internal consistency applies only to items in a measure; items in an index need not be correlated at all to be valid. Since distinguishing measures from indices can be difficult, strategies such as thought experiments supplemented by content validation studies using subject matter experts have been suggested (Zumbo, 2007a).

The second step in the framework is articulating the nature of the aggregate construct. First, it must be made clear at the outset that there is no such thing as a "correct" way to aggregate EDI scores. At any one level of aggregation (e.g., class, school, neighbourhood), several compositional models may be possible. Across different levels of aggregation, neither is it necessary to apply the same compositional model in each instance. The appropriateness of a particular compositional model depends on both one's multilevel theoretical expectations and observed patterns of within- and between-group variation.

Having said that, four of the compositional models described by Chen et al. (2004a) do not apply to the EDI (assuming that we are aggregating *individual scores* to some higher level, and not, for example, classroom-level scores to the school district level). These are the selected score, consensus, referent shift consensus, and aggregate properties models. The selected score model does not apply because it seems unlikely that any one individual's EDI score should be weighted any differently than any other individual's score when creating a group-level score. Consensus and referent-shift consensus models only apply when there is an expectation of agreement between individuals, and this is not the case for developmental measures like the EDI. Aggregate models do not apply because they are based on global scores, and it seems likely that aggregate EDI constructs will always be based on scores at the individual child level. As Hofmann and Jones (2004) would put it, we are concerned with summarizing individual EDI scores, not measuring something like the "emotional maturity of the neighbourhood."

In essentially all EDI analytic studies so far, whether the level of aggregation was classes, schools, school districts, or neighbourhoods, one of two methods have been used to create collective EDI scores. These aggregate scores have either been the mean of individual scores (e.g., La Pointe, 2006), or the percentage of "vulnerable" children based on established vulnerability cutoff scores for each developmental scale (e.g., Kershaw & Forer, 2006).

Aggregating EDI scores based on the group (e.g., class) mean is a clear example of a summary index compositional model. The aggregate level scores in this model are a function of the absolute values of the individual level scores. In contrast, the "percent vulnerable" aggregate scores do not fit any of Chen et al.'s (2004) model types exactly. As Kim (2004) observed, Chen's models do not account for aggregate scores based on within-group variability. "Percent vulnerable" has just this basis as it derives its meaning relative to a group-level referent. In the case of the EDI, however, the reference group is the whole province, and so all classes have the

same referent (i.e., the provincial 10% cutoff score in Wave 1). Therefore, the "percent vulnerable" score for a group has elements of both a summary index score (since the proportion for a binary variable is a sort of mean score) as well as an undefined composition model based on within-group variability (since binary vulnerability scores for individuals have been relatively defined).

In the following empirical analyses, the validation methods chosen assume a summary index composition model for both group means and "percent vulnerable" scores. The justification for the latter is that even though the cutoff score was derived in a relative way, it is being treated as permanent for the scoring of EDI vulnerability scores for all future waves of data collection. This reification transforms what started as a relative score into something more absolute – i.e., a score that is not expected to change as a result of an absolute (i.e., clinical) cutoff being determined.

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 a consideration of within-group agreement on item scores, factor structure across levels, and reliability of item scores. Within-group agreement is not relevant, as aggregate EDI scores are not based on models that rely on consensus. However, factor structure and reliability are both quite relevant.

Although Janus and Offord (2007) do not have an explicit theory to guide EDI score aggregation, the multilevel CFA technique they used does make the assumption of a summary index compositional model, using between-class mean EDI scores at the class level. This two-level CFA assumed unidimensionality for each EDI scale at each level. Janus and Offord (2007) expected to find (and did claim to find) factor structure similarity across the individual and class levels for the EDI, and interpreted their results as confirmation of an isomorphic multilevel

factor structure. The theory behind this expectation is unclear, but is consistent with their interpretation that between-class variation is relatively unimportant.

There can be issues of unreliability of parameter estimates (i.e., factor loadings, variances, and covariances) in multilevel CFA when sample sizes are small at aggregated levels. However, this does not apply to the EDI, given the large number of schools, neighbourhoods and school districts in the database.

Depending on the compositional model, reliability of the item scores at the group level can be calculated quite differently. This is because different models make different assumptions about systematic and error variance (Chen et al., 2004a). In consensus models, variation between individuals is considered to be error variance, while in models that describe a collection of individuals (including the summary index model), that same variation is considered to be true variance. When EDI scores are aggregated using group means, internal consistency reliability need only be calculated at the individual level. This type of reliability at the individual level is not meaningful for vulnerability scores, as vulnerability does not exist at the item level.

The fourth step in the multilevel construct validation process is an analysis of the relative amounts of within-group and between-group variation, which provides empirical guidance about appropriate levels of aggregation. Bliese (2000) and Chen et al. (2004a) discuss three different measures that can help assess whether data collected from individuals have group-level properties. First, the level of non-independence in data can be measured using an intraclass correlation coefficient called 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 within-group reliability, which is measured using a different intraclass correlation coefficient called ICC(2).This can be considered as a coefficient of generalizability (O'Brien, 1990); it indexes the reliability of differences between group means by measuring the

proportional consistency of variance across groups (Bliese, 2000; Ostroff, 1992). High within-group reliability means that group means can be estimated from a relatively small number of individuals in a group.

The third aspect of group properties is within-group agreement, the extent to which individuals within a group have the same absolute ratings. High within-group agreement provides justification for the aggregation of scores to the group level. It is most commonly measured using the r_{wg} statistic (James, Demaree, & Wolf, 1984), and can be low even when within-group reliability is high because the former assesses absolute consistency while the latter assesses relative consistency. Within-group agreement, as noted before, applies to the two consensus compositional models, but not to the selected score, summary index, or dispersion models. Thus, there is no requirement for EDI scores to have high within-group agreement as a prerequisite for appropriate aggregation. In contrast, high within-group reliability would be desirable, at least for group EDI scores based on the mean. This is because having reliable group differences makes it possible to model predicted group scores.

3.1.2 WABA Framework

Of the four inferential components in the WABA approach to validation (Dansereau et al., 1984), two apply to the same structural aspects of multilevel construct validation addressed by Chen et al. (2004a). The first is single level analysis (SLA), the purpose of which is to detect whether individual EDI scores indicate a wholes, parts, or equivocal situation when aggregated to the classroom level. As described in the literature review, this is accomplished by first testing practical significance based on E-ratio values, and following up with an F-ratio test of statistical significance.

The second inferential component is multiple level analysis (MLA), the purpose of which is to detect what happens to the EDI when aggregated to two or more nested levels – whether the single-level analysis inference (i.e., wholes, parts, equivocal) applies across both aggregate levels (cross-level), or if there are different inferences at each level of aggregation (level-specific or emergent). Unfortunately, in the case of the EDI, the two levels of aggregation cannot be the class and neighbourhood levels because classes are not consistently nested within neighbourhoods; it is common for children within the same classroom to live in different neighbourhoods. In fact, for the kindergarten classes in the study sample with at least five children, only 19.4% were composed of children who all live in the same neighbourhood. The ability to conduct WABA MLA analyses, however, is possible using the class and school district levels instead, given that all children in a particular classroom live in the same school district.

3.2 Methods – Applying Frameworks to EDI

3.2.1 Psychometric Properties: Multilevel Factor Structure

3.2.1.1 Analytic Issues and Potential Strategies

There are a number of similar strategies that have been suggested for conducting multilevel covariance structure (i.e., multilevel SEM) analyses (e.g., Hox, 2002; Muthen, 1994). Each strategy consists of a series of steps designed to assess within-group and between-group factor structure, first separately and later simultaneously. For example, there are five steps in Muthen's (1994) strategy: a) conducting a conventional factor analysis of overall covariance ($\mathbf{S_T}$) matrix, b) using the intraclass correlations to assess between-group variation, c) estimating the within-group structure using the pooled within-group covariance matrix, d) estimating the between-group structure using the sigma-between covariance matrix, and finally e) conducting the actual multilevel SEM analysis. At each step, the absolute fit of the model is assessed using a

variety of fit statistics such as the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), and the Standardized Root Mean Square Residual (SRMR). Relative differences in model fit from step to step are also assessed, typically using a chi-square difference test.

Unfortunately, for the EDI, due to computational considerations, it is not possible for Mplus 4.1 (the modeling software used) to produce the pooled within-group or sigma-between covariance matrices that are necessary for executing the stepwise Muthen (1994) strategy. For two-level SEM analyses with binary or ordered categorical dependent variables, as is the case for each EDI scale, only various forms of maximum likelihood (ML) estimation are allowed (Muthen & Muthen, 2006). These models are very computationally demanding, given that ML employs an algorithm where each item requires an additional dimension of mathematical function to be integrated. As a result, Mplus 4.1 cannot produce either of these covariance matrices or the usual model fit statistics (i.e., RMSEA, CFI, SRMR) for such models (L. Muthen, personal communication, March 16, 2007). The only fit statistics that are currently reported when ML estimation is used are the loglikelihood values (with an associated scaling correction factor), and Information Criteria statistics (Akaike and Bayesian).

This leaves few computationally viable statistical options at present for estimating the fit of multilevel models with categorical items. Grilli and Rampichini (2007) have recently adapted Muthen's (1994) stepwise approach for the categorical situation using a five-item scale, and dealing with the computational demands by making assumptions at each step. For example, their first step is to estimate the intraclass correlation coefficient (ICC[1]) for each item to assess the appropriateness of later multilevel factor analyses. They avoid problems with model underidentification by fixing the means and the individual-level standard deviations. Similarly, their third step involves conducting separate exploratory factor analyses using the between-group

and within-group correlation matrices, respectively. This presupposes the ability to decompose the total correlation matrix into the between and within components, which quickly becomes impossibly burdensome computationally when there are more than three or four of these categorical items (L. Muthen, personal communication, March 16, 2007). To reduce the burden, they recommend treating the items as continuous, with the additional assumption that the distribution of each item is Gaussian. Despite its promise, the Grilli and Rampichini (2007) strategy does not seem viable for the EDI, due to a) the large number (n=103) of items, and b) the categorical and non-Gaussian nature of these items.

Another approach would be to work with nested models, and test the relative fit of the models using a scaled chi-square difference test. For multilevel models with categorical items, the simple difference between the scaled chi-square statistics is inappropriate for this purpose, however, as it is not distributed as chi-square (Muthen & Muthen, 2006). Satorra and Bentler (1999) have calculated a chi-square difference test scaling correction factor to overcome this problem. This approach is not useful for evaluating the overall fit of EDI multilevel models without covariates, as there are no alternative models against which to test. However, this relative fit approach will be a useful analytic strategy for the second research question, which is to determine whether the fit of the multilevel model (for individuals and classes) for each scale is significantly improved by including available teacher/classroom covariates.

3.2.1.2 Strategy Employed For EDI Multilevel Factor Analysis

The approach that was followed for evaluating multilevel model fit was to examine the pattern of bivariate residuals after fitting the model to the data. This is akin to the approach used for the SRMR index (Hu & Bentler, 1995). The advantage of this approach over the chi-square difference test is that it assesses absolute fit. However, because the data are categorical, the

residuals are proportions rather than covariances. Examining the residual proportions does provide information about their magnitude, but any interpretation regarding model fit is necessarily subjective. Since these residuals are not standardized, the usual statistical inferences based on known distributional properties are not possible. In the results that follow, the magnitude of these residual proportions are reported for descriptive purposes, and a newly created version of a residuals-based fit test using proportions is described and applied to these residuals.

When data are continuous, there are $p(p-1)/2$ residuals, where p represents the number of items in the dataset. However, with categorical data, the overall number of residuals depends on the number of categories for each item as well as the number of items. For example, for each pair of binary items there are four combinations of categories, and therefore four bivariate residuals. Similarly, for each pair of three-category items, there are nine bivariate residual proportions. Therefore, for the whole social competence scale, which consists of 26 three-category items, there are 9*(26)*(26-1)/2 or 2,925 bivariate residual proportions.

3.2.1.3 Developing the Root Mean Square Residual – Proportion (RMR-P) Fit Index

While examining the residual proportions provides information about their magnitude and the effect of clustering on their distribution, any interpretation regarding fit is necessarily subjective. Since these residuals are not standardized, the usual criteria for statistical inferences based on known distributional properties are not possible. There is no objective basis upon which to make any conclusions about model fit, as there is no readymade fit index for residual proportions. Accordingly, a new fit index was developed based on the concept of the Root Mean Square Residual (RMR; Hu & Bentler, 1995), a commonly reported fit statistic for items measured on a continuous scale. The RMR is the average residual value over all residuals after

fitting the model covariance matrix to the data covariance matrix. To aid interpretation, these residuals are normally standardized by using the corresponding correlation matrices, resulting in a range from 0.0 to 1.0 for what is then called the Standardized Root Mean Square Residual (SRMR). An SRMR of zero indicates a perfectly fitting model, and as a rule of thumb, a model with an SRMR value less than .05 is deemed to have good fit.

The new Root Mean Square Residual-Proportion (RMR-P) fit statistic is calculated in a way that is similar to the SRMR. Details about the steps in this calculation can be found in Appendix A. Like the SRMR, the RMR-P has a maximum range from 0.0 to 1.0, with a value of .05 set as the cutoff for good fit.

3.2.2 Psychometric Properties - Dimensionality of EDI Scales at the Classroom and Neighbourhood Levels

The first multilevel SEM models of unidimensional EDI scales examine the influence of clustering on model fit for individual children. However, these results do not address the dimensionality of the EDI scales at either the classroom or neighbourhood level. These one-level analyses are an important aspect of the validation of the EDI, as the instrument was developed specifically to be interpreted using aggregated scores.

To assess the dimensionality of each EDI scale at these aggregated levels, a three-step analytic approach was employed. First, at each of the class and neighbourhood levels, a one-level CFA was conducted for each scale. The purpose of this step was to conduct a strict test of unidimensionality for each scale at each level, based on appropriate fit statistics. For those scales that did not meet this more exacting standard for unidimensionality, follow-up exploratory factor analyses (EFAs) were conducted as the second and third steps in the analysis. This second step involved conducting an eigenvalue ratio test of the EFA results (Gorsuch, 1983); this provided a

more liberal test of unidimensionality than the previous CFA analysis. If a scale (at a particular level) did not meet either of these two standards of unidimensionality, then the EFA proceeded so as to determine the number of factors necessary to have a combination of good model fit and interpretability.

The one-level CFA analyses reported below are based on the mean score per item, aggregated at the class or neighbourhood levels. Therefore, unlike the previous multilevel CFA analyses, the measures are continuous rather than categorical; this allows fit statistics to be calculated. A robust variant of maximum likelihood estimation was used, to be consistent with the previous multilevel analyses.

3.2.3 Psychometric Properties – Within-Group Agreement

Within-group agreement refers to the degree to which individuals in a group are exchangeable. It is most commonly measured using the intermember agreement index r_{wg}, developed by James et al. (1984). This index assesses whether individual variability in scores within each group is less than would be expected from a random distribution. There are variations of the r_{wg} index that employ different theoretical random distributions. Typically, an r_{wg} score of.70 or higher is considered high enough to justify meaningful group-level scoring. In the case of the EDI (and for non-consensus compositional models in general), there is no expectation of high agreement between individuals on their developmental domain scores, and so this index was not calculated.

3.2.4 Variability Between and Within Groups

The fourth step in the multilevel validation framework of Chen et al., (2004) is to examine the relative amounts of within-group and between-group variation, to assess whether individual-level data have group-level properties. Various forms of the intraclass correlation coefficient are used in this assessment. The first is called ICC(1), and indicates the proportion of the total variance in the scores that can be attributed to the aggregate level (i.e., classes or neighbourhoods in the context of this study). There are two different forms of the ICC(1), one based on a random-coefficient model (such as HLM) and the other based on a one-way random effects ANOVA. While both are interpreted similarly, the former ranges from 0 to +1, while the latter ranges from -1 to +1. In the current study, the random-effects ANOVA version of the ICC(1) was calculated.

3.3 Results

3.3.1 Multilevel Factor Structure, Actual Scores

Table 11 below summarizes the distributional properties for the residual proportions for each of the five scales, based on EDI actual item scores. Three different factor structure models (one-level, two-level clustered by class, and two-level clustered by neighbourhood) were estimated for each scale. For each model, the table shows the mean residual (which is always zero by definition), the standard deviation, the 5^{th} and 95^{th} percentiles, and the minimum and maximum residual values. Given that the mean residual is zero, the dispersion statistics are the best measures of how well each model fits the data.

The results in Table 11 provide information about both the magnitude of the residuals, and the differences in magnitude when clustering is included in the models. With regards to absolute magnitude, all models had quite modest residuals, subjectively speaking. The difference between the 5^{th} and 95^{th} percentile scores provides one measure of magnitude. For all 15 models, the table shows that 90% of the residual proportions were within .035 of the observed proportions. Even when considering only the most extreme (i.e., minimum and maximum) residual proportions, the magnitudes were not much larger. Only one scale, emotional maturity, had any residual proportions greater than .10 in magnitude, and for the majority of models, there were no residual proportions greater than .06.

Table 11 also shows how the distribution of residuals for each scale was affected by clustering. One would expect, if class or neighbourhood has an important influence, that taking clustering into account when estimating the factor structure should improve model fit (i.e., by reducing the dispersion of residual proportions). Potential reductions were examined based on the range and standard deviation of residual proportions when clustering is included. It should be kept in mind, however, that with residuals of such small absolute magnitude, there is relatively little scope for improvement.

Table 11 Residual Bivariate Proportions Based on Actual Scores, One-level and Two-level EDI models, by EDI scale

	Mean	*Standard Deviation*	*5th Percentile*	*95th Percentile*	*Minimum*	*Maximum*
Physical – 1 level	0.000	0.012	-0.018	0.019	-0.051	0.073
Physical – 2 levels, class	0.000	0.012	-0.015	0.021	-0.049	0.070
Physical – 2 levels, n'hood	0.000	0.010	-0.013	0.015	-0.049	0.067
Social – 1 level	0.000	0.010	-0.016	0.015	-0.085	0.091
Social – 2 levels, class	0.000	0.009	-0.014	0.014	-0.061	0.083
Social – 2 levels, n'hood	0.000	0.010	-0.016	0.016	-0.076	0.085
Emotional – 1 level	0.000	0.016	-0.020	0.025	-0.084	0.125
Emotional – 2 levels, class	0.000	0.012	-0.014	0.016	-0.052	0.104
Emotional – 2 levels, n'hood	0.000	0.015	-0.025	0.024	-0.082	0.123
Language – 1 level	0.000	0.007	-0.011	0.009	-0.043	0.044
Language – 2 levels, class	0.000	0.006	-0.009	0.007	-0.039	0.040
Language – 2 levels, n'hood	0.000	0.006	-0.009	0.008	-0.040	0.043
Comm. – 1 level	0.000	0.015	-0.022	0.032	-0.036	0.057
Comm. – 2 levels, class	0.000	0.015	-0.023	0.031	-0.037	0.056
Comm. – 2 levels, n'hood	0.000	0.015	-0.026	0.035	-0.037	0.059

Table 12 below shows the absolute and percent change in the range of residuals after clustering was included in the models. Absolute change was greater than .008 for only three of the 10 models, and in only two of those (social competence and emotional maturity at the classroom level) was the corresponding percentage reduction in range greater than 10%. These patterns are approximately repeated when looking at the reduction in the standard deviations of the residual proportions (see Table 11). Thus, from a purely descriptive standpoint, the results show that the already-small residual proportions found for the individual level models are not substantially reduced, for any EDI scale, by accounting for neighbourhood clustering. There

were two scales with larger reductions when classroom clustering is included, but the absolute changes are still relatively small.

Table 12 Absolute and Percent Change in Range of Residual Proportions, After Clustering by Class and Neighbourhood

	Class		***Neighbourhood***	
	Absolute change	***Percent change***	***Absolute change***	***Percent change***
Physical	-.005	-4	-.008	-6.5
Social	-.032	-18	-.015	-8.5
Emotional	-.053	-25	-.004	-2
Language	-.008	-9	-.004	-4.5
Communication	.000	0	.003	+3

3.3.1.1 RMR-P for EDI Actual Scores

The results for the overall RMR-P for each scale and across levels based on actual EDI scores can be found in Table 13. All five scales meet the good-fit criterion of an RMR-P less than .05, for the individual level model as well as both two-level models. Relatively speaking, the best model fit is for the language and cognitive development scale and the worst model fit is for the emotional maturity scale. For all scales, the RMR-P scores confirm that taking clustering into account (both class and neighbourhood levels) results in modest improvements in fit at best.

Table 13 Root Mean Square Residual – Proportion (RMR-P), One- and Two-level EDI Models, Based on Actual Scores, by EDI scale

	RMR-P
Physical – 1 level, individual	.018
Physical – 2 levels, individual and class	.018
Physical – 2 levels, individual and neighbourhood	.016
Social – 1 level, individual	.022
Social – 2 levels, individual and class	.019
Social – 2 levels, individual and neighbourhood	.021
Emotional – 1 level, individual	.034
Emotional – 2 levels, individual and class	.026
Emotional – 2 levels, individual and neighbourhood	.032
Language – 1 level, individual	.010
Language – 2 levels, individual and class	.008
Language – 2 levels, individual and neighbourhood	.009
Communication – 1 level, individual	.031
Communication – 2 levels, individual and class	.031
Communication – 2 levels, individual and neighbourhood	.032

Note: See Appendix A for the steps in calculating the RMR-P

Taken together, the above one-level and multilevel CFA analyses indicate a good-fitting unidimensional model for each scale at the level of individual students, with no notable improvements in the individual-level unidimensional measurement model by taking clustering into account. Therefore, these results seem to show that scale scores are meaningful for individual children, with no additional meaning required for aggregated scores.

3.3.2 Multilevel Factor Structure and RMR-P for EDI Vulnerability Scores

Given that the vulnerability scores are binary, a goodness-of-fit analysis of their multilevel factor structure must also focus on residual proportions rather than correlations or covariances. The main difference in this analysis is that there is only one CFA model over all five scales, rather than a separate model for each scale as there was for the item scores. This is because there are no item-level measures of vulnerability. Thus, the multilevel CFA in this case tests how well a model measuring a latent variable called “overall vulnerability” fits the data from five domains of vulnerability, and how this fit is affected by clustering at the classroom and

neighbourhood levels. The RMR-P fit statistic is used, as before, to provide an objective method for assessing fit.

Table 14 summarizes the same distribution properties of residual proportions as previously in Table 11, but in this instance for vulnerability scores. As before, the mean residual proportion is, by definition, equal to zero for all models. The standard deviation, 5^{th} and 95^{th} percentile scores, and maximum and minimum all summarize the absolute magnitude of the residuals and how the magnitude is influenced by clustering.

The table shows that the residual proportions are even smaller than was the case for the actual scores. All three models fit the data very well, and there is no discernible improvement in fit by accounting for clustering at either the class or neighbourhood levels. As with the actual scores at the scale level, the multilevel factor structure seems to suggest that vulnerability is meaningful at the individual level, with no requirement for a different interpretation when aggregated.

Table 14 Residual Bivariate Proportions Based on Vulnerability Scores, One-level and Two-level EDI Models

	Mean	*Standard Deviation*	*5^{th} Percentile*	*95^{th} Percentile*	*Minimum*	*Maximum*	*RMR-P*
One level (ind. only)	0.000	0.006	-0.013	0.008	-0.013	0.014	.009
Two levels (ind. & class)	0.000	0.006	-0.013	0.009	-0.014	0.017	.009
Two levels (ind. & n'hood)	0.000	0.007	-0.014	0.010	-0.015	0.015	.010

3.3.3 Dimensionality of the EDI scales at Aggregated Levels

So far, the results have shown that multilevel SEM models of unidimensional EDI scales fit well for scores for individual children. However, these results do not address the dimensionality of the EDI scales at either the classroom or neighbourhood level. These one-level analyses are an important aspect of the validation of the EDI, as the instrument was developed specifically to be interpreted using aggregated scores.

Table 15 shows three model fit statistics for each scale, at both the class and neighbourhood levels. The Root Mean Square Error of Approximation (RMSEA) is considered to be the most informative fit statistic (Byrne, 1998), and so was given primary consideration. RMSEA values below .05 are indicative of good fit, though values as high as .08 are considered adequate. Based on this criterion, there were only two instances where the items adequately fit a one-dimensional model of the scale – physical health and well-being at the class level, and communication and general knowledge at the neighbourhood level. However, for the former, the RMSEA was exactly at the .08 adequate-fit threshold, and the two other fit statistics did not reach the criteria for good fit (i.e., >.90 for CFI and <.05 for SRMR). Thus, there is justification for further investigation using exploratory factor analysis (EFA). For the communication and general knowledge scale at the neighbourhood level, however, model fit was good for all three fit statistics based on the above criteria, and so no further analytic steps are needed before declaring this to be an instance of unidimensionality.

Table 15 Fit Statistics – One-level CFAs for each EDI scale, Class and Neighbourhood Levels

	CFI	***RMSEA***	***SRMR***
Physical – class level	0.793	**0.080**	0.104
Physical – neighbourhood level	0.700	0.127	0.125
Social – class level	0.732	0.113	0.077
Social – neighbourhood level	0.639	0.106	0.097
Emotional – class level	0.475	0.118	0.148
Emotional – neighbourhood level	0.403	0.119	0.149
Language – class level	0.681	0.087	0.080
Language – neighbourhood level	0.591	0.105	0.088
Communication – class level	**0.964**	0.097	**0.018**
Communication – neighbourhood level	**0.966**	**0.054**	**0.025**

Notes. CFI = Comparative Fit Index
RMSEA = Root Mean Square Error of Approximation
SRMR = Standardized Root Mean Square Residual
Bolded values indicate adequate fit

3.3.3.1 Exploratory Factor Analyses at the Class and Neighbourhood Levels

For all scale/level combinations except Communication and General Knowledge at the neighbourhood level, follow up exploratory factor analyses (EFAs) were conducted in Mplus 4.1. To be consistent with the earlier CFA analyses, a robust form of maximum likelihood estimation called MLMV was chosen for these analyses. Solutions with one factor through to four factors were considered as this is the Mplus default and offered a reasonable number of solutions to examine. As mentioned previously, a two-step approach was used in the EFA analyses. The first step was to examine the ratio of the first and second eigenvalues, and the second, if necessary, was to examine the RMSEA fit statistic for each solution (i.e., different number of factors extracted) in combination with the simple-structure interpretability of the factor loadings after promax rotation.

3.3.3.2 Exploratory Factor Analysis: Eigenvalue Ratio Results

As a rule of thumb, an eigenvalue ratio of 3.0 or greater can be considered as evidence of essential unidimensionality (Gorsuch, 1983). The eigenvalue ratio for each scale at each level of aggregation is shown in Table 16. The results indicate essential unidimensionality for three of the five EDI scales (social competence, language and cognitive development, and communication and general knowledge), consistently at both levels of aggregation. As such, the final EFA step of exploring different factor solutions was only conducted for the physical health and well-being and emotional maturity scales.

Table 16 Exploratory Factor Analysis, Ratio of Eigenvalues (First:Second), All Scales, Class and Neighbourhood Levels

	Class level	***Neighbourhood level***
Physical Health & Well-being	2.46	2.40
Social Competence	7.00	5.13
Emotional Maturity	2.78	2.60
Language and Cognitive Development	4.97	5.40
Communication and General Knowledge	10.38	*

Note. * EFA not necessary, as CFA strict test already indicates unidimensionality

3.3.3.3 Exploratory Factor Analysis: RMSEA and Interpretability Results

As was the case for the one-level CFA analyses in Section 5.2.2.5, only models with RMSEA statistics of .08 or less were considered to have adequate fit. With respect to factor loadings, only loadings of .40 or larger were interpreted, as recommended by Stevens (2009).

For the physical health and well-being scale at the class level, the RMSEA for the one-factor solution is .074, indicative of adequate fit. However, seven of the 13 items are not interpretable for this factor, having loadings well under 0.40, as shown in Table 17. The two-factor solution has an improved RMSEA of 0.055, and better interpretability. The first factor has four items, the second has five items, and there are four items (A6 to A8, and C58) that are not associated with either factor (see Table 18). The three-factor solution (not shown), as expected, has an even lower RMSEA of 0.039, but the same four items are still not associated with any factor. Therefore, the two-factor solution seems best, even though not all items load strongly on either factor.

Table 17 Factor Loadings, One-Factor Solution, Physical Health and Well-Being Scale Items Aggregated to the Class Level

Item	Factor 1
A2. Over or underdressed for school-related activities	0.170
A3. Too tired/sick to do school work	0.249
A4. Late	0.145
A5. Hungry	0.208
A6. Independent in washroom activities most of the time	0.210
A7. Shows an established hand preference	0.203
A8. Is well-coordinated	**0.405**
A9. Proficiency at holding pen, crayons, or brush	**0.820**
A10. Ability to manipulate objects	**0.912**
A11. Ability to climb stairs	**0.892**
A12. Level of energy throughout the school day	**0.866**
A13. Overall physical development	**0.924**
C58. Sucks a thumb/finger	0.123

Note: Factor loadings greater than .40 are **bolded**

Table 18 Promax Rotated Factor Loadings, Two-Factor Solution, Physical Health and Well-Being Scale Items Aggregated to the Class Level

Item	Factor 1	Factor 2
A2. Over or underdressed for school-related activities	**0.579**	-0.070
A3. Too tired/sick to do school work	**0.748**	-0.057
A4. Late	**0.474**	-0.051
A5. Hungry	**0.624**	-0.049
A6. Independent in washroom activities most of the time	0.163	0.143
A7. Shows an established hand preference	0.176	0.128
A8. Is well-coordinated	0.364	0.255
A9. Proficiency at holding pen, crayons, or brush	0.006	**0.813**
A10. Ability to manipulate objects	-0.013	**0.915**
A11. Ability to climb stairs	-0.086	**0.932**
A12. Level of energy throughout the school day	0.117	**0.820**
A13. Overall physical development	-0.027	**0.941**
C58. Sucks a thumb/finger	0.213	0.035

Note: Factor loadings greater than .40 are **bolded**

At the neighbourhood level, the pattern for the physical health and well-being scale is similar. The fit of the one-factor solution is very good (RMSEA = 0.049), but, as Table 19 shows, six items do not load on that factor. Nine of the 13 items load 0.40 or higher on one of the factors in the two-factor solution (RMSEA = 0.054; see Table 20), as is also the case for the three-factor solution (not shown). Again, the two-factor solution is preferred.

Table 19 Factor Loadings, One-Factor Solution, Physical Health and Well-Being Scale Items Aggregated to the Neighbourhood Level

Item	Factor 1
A2. Over or underdressed for school-related activities	0.156
A3. Too tired/sick to do school work	0.313
A4. Late	0.135
A5. Hungry	0.326
A6. Independent in washroom activities most of the time	0.145
A7. Shows an established hand preference	0.227
A8. Is well-coordinated	**0.418**
A9. Proficiency at holding pen, crayons, or brush	**0.864**
A10. Ability to manipulate objects	**0.920**
A11. Ability to climb stairs	**0.915**
A12. Level of energy throughout the school day	**0.904**
A13. Overall physical development	**0.936**
C58. Sucks a thumb/finger	0.209

Note: Factor loadings greater than .40 are **bolded**

Table 20 Promax Rotated Factor Loadings, Two-Factor Solution, Physical Health and Well-Being Scale Items Aggregated to the Neighbourhood Level

Item	Factor 1	Factor 2
A2. Over or underdressed for school-related activities	**0.680**	-0.112
A3. Too tired/sick to do school work	**0.824**	-0.007
A4. Late	0.382	-0.016
A5. Hungry	**0.903**	-0.023
A6. Independent in washroom activities most of the time	0.072	0.117
A7. Shows an established hand preference	-0.031	0.240
A8. Is well-coordinated	**0.446**	0.244
A9. Proficiency at holding pen, crayons, or brush	0.006	**0.827**
A10. Ability to manipulate objects	0.096	**0.938**
A11. Ability to climb stairs	-0.058	**0.940**
A12. Level of energy throughout the school day	0.041	**0.888**
A13. Overall physical development	-0.034	**0.950**
C58. Sucks a thumb/finger	0.081	0.176

Note: Factor loadings greater than .40 are **bolded**

For the emotional maturity scale at the class level, fit is poor for the one-factor solution (RMSEA = 0.114), and adequate for the two-factor solution (RMSEA = 0.076). After promax rotation, the latter solution consists of an eight-item factor and a 20-item factor, with two items (C36 and C57) with loadings under 0.40 (see Table 21). Interpretability and fit are both a little

improved for the three-factor solution shown in Table 22. The RMSEA statistic is now 0.063, and there is now an eight-item factor, a 14-item factor, and a seven-item factor, with only item C36 not interpretable for any factor. Either the two-factor or three-factor solution is justified for this scale at the class level.

Table 21 Promax Rotated Factor Loadings, Two-Factor Solution, Emotional Maturity Scale Items Aggregated to the Class Level

Item	Factor 1	Factor 2
C28. Willing to try to help someone who is hurt	**0.847**	0.056
C29. Volunteers to help clear mess made by other	**0.781**	0.115
C30. Will try to stop quarrel or dispute	**0.892**	-0.003
C31. Offers to help others having task difficulty	**0.873**	0.057
C32. Comforts a child who is crying or upset	**0.956**	-0.045
C33. Spontaneously picks up others' dropped objects	**0.788**	0.076
C34. Invites bystanders to join in a game	**0.801**	0.014
C35. Helps other children who are feeling sick	**0.957**	-0.071
C36. Is upset when left by parent/guardian	-0.062	0.320
C37. Gets into physical fights	-0.063	**0.591**
C38. Bullies or is mean to others	-0.006	**0.624**
C39. Kicks, bites, hits other children or adults	-0.016	**0.624**
C40. Takes things that do not belong to him/her	-0.013	**0.580**
C41. Laughs at other children's discomfort	0.050	**0.488**
C42. Can't sit still, is restless	-0.006	**0.855**
C43. Is distractible, has trouble sticking to activities	0.014	**0.857**
C44. Fidgets	-0.008	**0.831**
C45. Is disobedient	0.049	**0.748**
C46. Has temper tantrums	0.006	**0.611**
C47. Is impulsive, acts without thinking	-0.005	**0.798**
C48. Has difficulty waiting turn in games or groups	-0.005	**0.794**
C49. Can't settle on anything for any length of time	0.000	**0.802**
C50. Is inattentive	0.077	**0.755**
C51. Seems to be unhappy, sad or depressed	0.026	**0.522**
C52. Appears fearful or anxious	-0.009	**0.490**
C53. Appears worried	-0.020	**0.460**
C54. Cries a lot	-0.031	**0.518**
C55. Is nervous, high-strung, or tense	-0.024	**0.543**
C56. Is incapable of making decisions	0.078	**0.541**
C57. Is shy	0.122	0.221

Note: Factor loadings greater than .40 are **bolded**

Table 22 Promax Rotated Factor Loadings, Three-Factor Solution, Emotional Maturity Scale Items Aggregated to the Class Level

Item	Factor 1	Factor 2	Factor 3
C28. Willing to try to help someone who is hurt	**0.849**	0.071	-0.030
C29. Volunteers to help clear mess made by other	**0.784**	0.137	-0.042
C30. Will try to stop quarrel or dispute	**0.894**	-0.031	0.038
C31. Offers to help others having task difficulty	**0.876**	0.053	-0.002
C32. Comforts a child who is crying or upset	**0.957**	-0.041	-0.011
C33. Spontaneously picks up others' dropped objects	**0.791**	0.089	-0.026
C34. Invites bystanders to join in a game	**0.804**	-0.016	0.037
C35. Helps other children who are feeling sick	**0.958**	-0.067	-0.014
C36. Is upset when left by parent/guardian	-0.064	0.096	0.354
C37. Gets into physical fights	-0.058	**0.620**	-0.054
C38. Bullies or is mean to others	-0.002	**0.620**	-0.002
C39. Kicks, bites, hits other children or adults	-0.012	**0.650**	-0.049
C40. Takes things that do not belong to him/her	-0.009	**0.601**	-0.041
C41. Laughs at other children's discomfort	0.052	**0.467**	0.031
C42. Can't sit still, is restless	-0.014	**0.916**	-0.065
C43. Is distractible, has trouble sticking to activities	0.009	**0.883**	-0.018
C44. Fidgets	-0.015	**0.891**	-0.063
C45. Is disobedient	0.052	**0.728**	0.025
C46. Has temper tantrums	0.009	**0.474**	0.204
C47. Is impulsive, acts without thinking	-0.003	**0.791**	0.010
C48. Has difficulty waiting turn in games or groups	-0.002	**0.736**	0.085
C49. Can't settle on anything for any length of time	0.000	**0.819**	-0.019
C50. Is inattentive	0.074	**0.743**	0.033
C51. Seems to be unhappy, sad or depressed	0.020	0.131	**0.625**
C52. Appears fearful or anxious	-0.025	-0.089	**0.952**
C53. Appears worried	-0.036	-0.126	**0.961**
C54. Cries a lot	-0.032	0.240	**0.436**
C55. Is nervous, high-strung, or tense	-0.033	0.135	**0.662**
C56. Is incapable of making decisions	0.075	0.260	**0.449**
C57. Is shy	0.119	-0.141	**0.576**

Note: Factor loadings greater than .40 are **bolded**

At the neighbourhood level for the emotional maturity scale, the one-factor solution has only mediocre fit (RMSEA = 0.086). The fit is improved (RMSEA = 0.054) for the two-factor solution (see Table 23), but three of the 30 items (C36, C40, and C57) have loadings under 0.40 for either factor. The three-factor solution (Table 24) has the best combination of fit (RMSEA = 0.046) and interpretability, as all 26 items load on one factor only.

Table 23 Promax Rotated Factor Loadings, Two-Factor Solution, Emotional Maturity Scale Items Aggregated to the Neighbourhood Level

Item	Factor 1	Factor 2
C28. Willing to try to help someone who is hurt	**0.878**	-0.004
C29. Volunteers to help clear mess made by other	**0.728**	0.148
C30. Will try to stop quarrel or dispute	**0.840**	0.018
C31. Offers to help others having task difficulty	**0.875**	0.037
C32. Comforts a child who is crying or upset	**0.982**	-0.050
C33. Spontaneously picks up others' dropped objects	**0.692**	0.121
C34. Invites bystanders to join in a game	**0.778**	0.115
C35. Helps other children who are feeling sick	**0.967**	-0.057
C36. Is upset when left by parent/guardian	-0.121	0.298
C37. Gets into physical fights	0.045	**0.554**
C38. Bullies or is mean to others	-0.006	**0.653**
C39. Kicks, bites, hits other children or adults	0.098	**0.539**
C40. Takes things that do not belong to him/her	0.171	0.382
C41. Laughs at other children's discomfort	0.060	**0.501**
C42. Can't sit still, is restless	0.123	**0.736**
C43. Is distractible, has trouble sticking to activities	0.109	**0.790**
C44. Fidgets	-0.019	**0.738**
C45. Is disobedient	0.056	**0.717**
C46. Has temper tantrums	-0.108	**0.585**
C47. Is impulsive, acts without thinking	0.010	**0.751**
C48. Has difficulty waiting turn in games or groups	0.016	**0.760**
C49. Can't settle on anything for any length of time	-0.084	**0.796**
C50. Is inattentive	0.074	**0.770**
C51. Seems to be unhappy, sad or depressed	0.015	**0.546**
C52. Appears fearful or anxious	-0.055	**0.499**
C53. Appears worried	-0.072	**0.464**
C54. Cries a lot	-0.169	**0.591**
C55. Is nervous, high-strung, or tense	-0.044	**0.540**
C56. Is incapable of making decisions	-0.041	**0.551**
C57. Is shy	0.079	0.171

Note: Factor loadings greater than .40 are **bolded**

Table 24 Promax Rotated Factor Loadings, Three-Factor Solution, Emotional Maturity Scale Items Aggregated to the Neighbourhood Level

Item	Factor 1	Factor 2	Factor 3
C28. Willing to try to help someone who is hurt	**0.867**	-0.112	0.073
C29. Volunteers to help clear mess made by other	**0.725**	0.012	0.145
C30. Will try to stop quarrel or dispute	**0.853**	0.108	-0.057
C31. Offers to help others having task difficulty	**0.881**	0.044	0.005
C32. Comforts a child who is crying or upset	**0.977**	-0.055	-0.013
C33. Spontaneously picks up others' dropped objects	**0.698**	0.086	0.066
C34. Invites bystanders to join in a game	**0.786**	0.089	0.056
C35. Helps other children who are feeling sick	**0.961**	-0.069	-0.011
C36. Is upset when left by parent/guardian	-0.092	**0.453**	0.011
C37. Gets into physical fights	0.005	-0.131	**0.686**
C38. Bullies or is mean to others	-0.042	-0.066	**0.740**
C39. Kicks, bites, hits other children or adults	0.053	-0.205	**0.717**
C40. Takes things that do not belong to him/her	0.155	-0.001	**0.405**
C41. Laughs at other children's discomfort	0.026	-0.104	**0.610**
C42. Can't sit still, is restless	0.113	0.090	**0.689**
C43. Is distractible, has trouble sticking to activities	0.108	0.198	**0.671**
C44. Fidgets	-0.017	0.197	**0.615**
C45. Is disobedient	0.032	0.070	**0.710**
C46. Has temper tantrums	-0.111	0.211	**0.471**
C47. Is impulsive, acts without thinking	-0.036	-0.130	**0.885**
C48. Has difficulty waiting turn in games or groups	-0.015	-0.027	**0.815**
C49. Can't settle on anything for any length of time	-0.097	0.145	**0.728**
C50. Is inattentive	0.060	0.083	**0.732**
C51. Seems to be unhappy, sad or depressed	0.055	**0.737**	0.093
C52. Appears fearful or anxious	-0.004	**0.882**	-0.047
C53. Appears worried	-0.017	**0.850**	-0.074
C54. Cries a lot	-0.140	**0.591**	0.228
C55. Is nervous, high-strung, or tense	-0.026	**0.405**	0.287
C56. Is incapable of making decisions	-0.013	**0.516**	0.229
C57. Is shy	0.137	**0.679**	-0.278

Note: Factor loadings greater than .40 are **bolded**

3.3.4 Psychometric Properties – Internal Consistency Reliability

Internal consistency reliability can be calculated using the actual EDI scores in each of the five scales, but it cannot be calculated for the vulnerability scores as they only exist at the scale level. According to Chen et al. (2004a), for summary index composition models, as is the case for the EDI actual scores, internal consistency should be assessed at the individual level only. Table 25 below shows the coefficient alpha values for each of the five EDI scales. All scales have internal consistency reliability higher than the commonly-held acceptable threshold of .70. The reliability values obtained are very similar to those reported by Janus, Walsh, and Duku (2005).

Table 25 Internal Consistency, Actual Scores, Individual Level, All Five EDI Scales

	Number of Students	*Coefficient alpha*
Physical Health and Well-Being	21,504	.782
Social Competence	21,163	.957
Emotional Maturity	18,185	.926
Language and Cognitive Development	20,472	.917
Communication and General Knowledge	24,774	.938

In the current study, the random-effects ANOVA version of the ICC(1) was calculated. As Table 26 below shows, ICC(1) values at the class level of aggregation using the actual EDI scores range from .19 to .25, depending on the EDI scale. In other words, from 19% to 25% of the variation in scores can be attributed to something about classrooms or teachers. These results are similar to those found by Janus and Offord (2007) for other Canadian kindergarten children, where the ICC(1) values for each scale ranged from .20 to .31. It is interesting that Janus and Offord (2007) interpret their ICC(1) values as "indicating low levels of variability between classrooms or teachers" (p. 10) due to most of the variability existing at the individual-child level. Many statisticians would consider ICC(1) values over .15 to be large (e.g., Hox, 2002).

Table 26 shows that the ICC(1) values for the vulnerability scores were substantially lower than those for the actual scores, for all scales and at both the class and neighbourhood levels. This makes sense, as teachers are doing the EDI rating directly, while the vulnerability scores are derived scores that are based on the provincial distribution of scores.

Neighbourhood level ICC(1) values range from .04 to .07 across scales, based on the actual scores. These values are much lower than the corresponding classroom level ICC(1) values.

Table 26 Intraclass Correlation Coefficient (1), by Scale, Level of Analysis, and Mean Scores vs. Vulnerability Scores

	Class		***Neighbourhood***	
	Mean	***Risk***	***Mean***	***Risk***
Physical Health and Well-Being	.236	.109	.063	.027
Social Competence	.185	.118	.040	.025
Emotional Maturity	.207	.096	.053	.023
Language and Cognitive Development	.220	.122	.071	.018
Communication and General Knowledge	.254	.150	.073	.043

All p's < .001

A second measure to assess the appropriateness of aggregation is the ICC(2) statistic, which some researchers label as ICC(1,k) or ICC(k) (Bliese, 2000). ICC(2) estimates the reliability of the group means, and is calculated using the following random-effects ANOVA-based formula: (MSB – MSW)/MSB. Thus, it ranges from undefined (when there is no variation between groups) to 1 (when there is some variation between groups, but no variation within groups). In keeping with the general rule of thumb for indices of reliability, ICC(2) values of .70 or greater are interpreted as indicating acceptable reliability, while values between .50 and .70 reflect marginal reliability, with poor reliability for ICC(2) values under .50. Table 27 shows that, the mean scores are reliable at both the classroom/teacher and neighbourhood levels, with values ranging from 0.72 to 0.84. However, the reliability of the corresponding vulnerability

scores are always lower than for the mean scores, typically only reaching a level of marginal reliability. Only the language and cognitive development scale at the neighbourhood level, and the communication and general knowledge scale at both levels had acceptable levels of ICC(2) reliability for the vulnerability scores.

Table 27 Intraclass Correlation Coefficient(2), by Scale, Level of Analysis, and Composition Model

	Class		***Neighbourhood***	
	Mean	***Risk***	***Mean***	***Risk***
Physical Health and Well-Being	.822	.646	.806	.626
Social Competence	.772	.667	.717	.610
Emotional Maturity	.794	.611	.773	.593
Language and Cognitive Development	.808	.674	.825	.697
Communication and General Knowledge	.836	.699	.830	.734

3.3.5 WABA Single Level Analyses (SLA)

The results of the WABA SLA, based on actual scores at the class level for each of the five EDI scales, is shown in Table 28 below. For all scales, the preponderance of total variation was within-classes, as shown by the relative values of the eta correlations within and between, as well as the corresponding E-ratios. The inference of "parts" is based on the practical significance criteria established for WABA. If the E-ratio is less than 0.767, then an inference of parts is made, based on the 15^{o} criterion. This is the inference for all EDI scales, based on the WABA analysis. When the E-ratio falls below 0.577, the more stringent 30^{o} criterion is exceeded, as was the case for the social competence scale only.

When a "parts" inference is made in terms of practical significance, an F-ratio test is used for statistical significance, but in this case, the ratio is reversed from the usual one-way ANOVA F-ratio. As Table 28 shows, the F value is less than one for all scales, and so the inference of a within-class effect is rejected. In fact, the more usual F-ratio that tests for a between-class effect is strongly significant for all scales. When this pattern of results is found, WABA analysts

(Dansereau, pers. comm.) conclude that an "equivocal" inference is appropriate (i.e., that there is no consistent evidence of a group effect, either between- or within-groups). Thus, they would recommend against using mean or total EDI scores aggregated at the class level.

Table 28 Results of the WABA Single Level Analysis Using Actual Scores, Class Level, All Five EDI Scales

	Phy	*Soc*	*Emo*	*Lan*	*Com*
Number of students	21,504	21,163	18,185	20,472	24,774
Number of classes	1,555	1,688	1,508	1,628	1,692
Eta correlation between	0.542	0.494	0.537	0.546	0.559
Eta correlation within	0.841	0.870	0.844	0.838	0.829
E-ratio	0.644	0.567	0.636	0.651	0.674
E-ratio Inference	Parts 15^{o}	Parts 30^{o}	Parts 15^{o}	Parts 15^{o}	Parts 15^{o}
F-ratio 1 (MS_{With}/MS_{Bet})	0.19	0.23	0.22	0.20	0.16
Probability for F-ratio 1	1.00	1.00	1.00	1.00	1.00
F-ratio 2 (MS_{Bet}/ MS_{With})	5.44	4.30	4.58	4.84	6.12
Probability for F-ratio 2	<.001	<.001	<.001	<.001	<.001
Final Inference	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal

Table 29 repeats the same WABA analysis at the class level, this time using EDI vulnerability scores for each scale, rather than actual scores. Unlike the actual scores, vulnerability scores are binary. Each child is either vulnerable or not for a particular scale, based on a comparison with children throughout the province. The results show the same patterns and inferences as above, despite the relatively lower E-ratios for the vulnerability scores for all scales.

Table 29 Results of the WABA Single Level Analysis Using Vulnerability Scores, Class Level, All Five EDI Scales

	Phy	*Soc*	*Emo*	*Lan*	*Com*
Number of students	21,504	21,163	18,185	20,472	24,774
Number of classes	1,555	1,688	1,508	1,628	1,692
Eta correlation between	0.419	0.425	0.423	0.450	0.444
Eta correlation within	0.908	0.905	0.906	0.893	0.896
E-ratio	0.461	0.470	0.467	0.504	0.495
E-ratio Inference	Parts 30^{o}	Parts 30^{o}	Parts 30^{o}	Parts 30^{o}	Parts 30^{o}
F-ratio 1 (MS_{With}/MS_{Bet})	0.37	0.34	0.41	0.34	0.30
Probability for F-ratio 1	1.00	1.00	1.00	1.00	1.00
F-ratio 2 (MS_{Bet}/ MS_{With})	2.73	2.94	2.42	2.93	3.34
Probability for F-ratio 2	<.001	<.001	<.001	<.001	<.001
Final Inference	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal

Tables 30 and 31 repeat the same WABA analyses, but for scores (actual and vulnerability, respectively) grouped at the neighbourhood level rather than the class level. The same two general patterns hold: 1) a "parts" inference based on the E-ratio, a "wholes" inference based on the F-ratio, resulting in an overall "equivocal" inference, and 2) lower E-ratios for vulnerability scores than for actual scores. Compared to the two respective class-level WABA analyses above, E-ratios at the neighbourhood level are consistently smaller, reflecting a greater proportion of within-group variation. This pattern of results relates to the larger group sizes at the neighbourhood level of aggregation, which are in the range of 45 to 60 children (compared to 12 to 15 at the class level). Generally, as group size increases, so does the proportion of total variation that is within-group.

Table 30 Results of the WABA Single Level Analysis Using Actual Scores, Neighbourhood Level, All Five EDI Scales

	Phy	*Soc*	*Emo*	*Lan*	*Com*
Number: students	21,394	24,044	18,095	20,371	24,652
Number: neighbourhoods	409	411	406	406	411
Eta correlation between	0.289	0.237	0.287	0.320	0.301
Eta correlation within	0.957	0.971	0.958	0.948	0.954
E-ratio	0.302	0.244	0.299	0.337	0.316
E-ratio Inference	Parts 30^o	Parts 30^o	Parts 30^o	Parts 30^o	Parts 30^o
F-ratio 1 (MS_{With}/MS_{Bet})	0.21	0.29	0.26	0.18	0.17
Probability for F-ratio 1	1.00	1.00	1.00	1.00	1.00
F-ratio 2 (MS_{Bet}/ MS_{With})	4.78	3.44	3.99	5.53	5.87
Probability for F-ratio 2	<.001	<.001	<.001	<.001	<.001
Final Inference	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal

Table 31 Results of the WABA Single Level Analysis Using Vulnerability Scores, Neighbourhood Level, All Five EDI Scales

	Phy	*Soc*	*Emo*	*Lan*	*Com*
Number: students	21,394	24,044	18,095	20,371	24,652
Number: neighbourhoods	409	411	406	406	411
Eta correlation between	0.216	0.205	0.218	0.250	0.247
Eta correlation within	0.976	0.979	0.976	0.968	0.969
E-ratio	0.221	0.210	0.223	0.258	0.255
E-ratio Inference	Parts 30^o	Parts 30^o	Parts 30^o	Parts 30^o	Parts 30^o
F-ratio 1 (MS_{With}/MS_{Bet})	0.39	0.39	0.46	0.31	0.26
Probability for F-ratio 1	1.00	1.00	1.00	1.00	1.00
F-ratio 2 (MS_{Bet}/ MS_{With})	2.51	2.54	2.18	3.28	3.85
Probability for F-ratio 2	<.001	<.001	<.001	<.001	<.001
Final Inference	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal

3.3.6 WABA Multiple-level Analyses (MLA)

Given that the SLA analyses above showed an "equivocal" inference at the class level, there are three potential multilevel patterns of EDI scores that could be revealed when taking the higher level school district into account. Either classes within a school district are homogeneous ("wholes" situation), interdependent ("parts" situation), or independent ("equivocal" situation). The first two possibilities are examples of what Dansereau et al. (1984) would label "emergent"

conditions, where the influences of grouping do not become apparent until the more collective level of aggregation is examined. The results of MLA analyses for both actual and vulnerability scores are shown below.

Table 32 summarizes the MLA analyses for each scale using actual EDI scores. The top half of the table repeats the SLA class-level results found earlier, while the bottom half of the table shows the results of considering the school district level. The familiar pattern of a "parts" inference using the E-ratio, and a "wholes" inference using the F-ratio is found, and so the final inference at the school district level is "equivocal." Since both levels of aggregation have the same "equivocal" inference, the overall MLA inference is also "equivocal." In other words, there is no emergent group effect (either "parts" or "wholes") found by taking the higher level of aggregation into account. EDI scores remain meaningful at the individual level only, according to the WABA MLA.

Table 32 Results of the WABA Multiple Level Analysis Using Actual Scores, Class and School District Levels, All Five EDI Scales

	Phy	*Soc*	*Emo*	*Lan*	*Com*
Number: students	21,394	24,044	18,095	20,371	24,652
Number: classes	1,555	1,688	1,508	1,628	1,692
Number: school districts	44	44	44	44	44
Class level					
Eta correlation between	0.542	0.494	0.537	0.546	0.559
Eta correlation within	0.841	0.870	0.844	0.838	0.829
E-ratio	0.644	0.567	0.636	0.651	0.674
E-ratio Inference	Parts 15^o	Parts 30^o	Parts 15^o	Parts 15^o	Parts 15^o
F-ratio 1 (MS_{With}/MS_{Bet})	0.19	0.23	0.22	0.20	0.16
Probability for F-ratio 1	1.00	1.00	1.00	1.00	1.00
F-ratio 2 (MS_{Bet}/ MS_{With})	5.44	4.30	4.58	4.84	6.12
Probability for F-ratio 2	<.001	<.001	<.001	<.001	<.001
Inference, Class level	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal
School District level					
Eta correlation between	0.258	0.234	0.252	0.345	0.246
Eta correlation within	0.966	0.972	0.968	0.938	0.969
E-ratio	0.267	0.241	0.261	0.368	0.254
E-ratio Inference	Parts 30^o	Parts 30^o	Parts 30^o	Parts 30^o	Parts 30^o
F-ratio 1 (MS_{With}/MS_{Bet})	0.40	0.45	0.43	0.20	0.40
Probability for F-ratio 1	1.00	1.00	1.00	1.00	1.00
F-ratio 2 (MS_{Bet}/ MS_{With})	2.51	2.21	2.32	4.99	2.47
Probability for F-ratio 2	<.001	<.001	<.001	<.001	<.001
Inference, District level	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal
Overall MLA Inference	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal

In Table 33, the same multiple-level analysis is conducted, but this time using vulnerability scores for each scale. The same pattern of results and conclusions found above also pertain to this MLA. The only slight difference is that at school district level for the emotional maturity scale, the traditional F-ratio did not quite achieve statistical significance. This does not change the conclusion of an "equivocal" situation, given that there is no statistical support for the "parts" inference implied by the E-ratio. It is interesting, however, that the F-ratio for this particular scale is different than for the other four scales.

Table 33 Results of the WABA Multiple Level Analysis Using Vulnerability Scores, Class and School District Levels, All Five EDI Scales

	Phy	*Soc*	*Emo*	*Lan*	*Com*
Number: students	21,394	24,044	18,095	20,371	24,652
Number: classes	1,555	1,688	1,508	1,628	1,692
Number: school districts	44	44	44	44	44
Class level					
Eta correlation between	0.419	0.425	0.423	0.450	0.444
Eta correlation within	0.908	0.905	0.906	0.893	0.896
E-ratio	0.461	0.470	0.467	0.504	0.495
E-ratio Inference	Parts 30^{o}	Parts 30^{o}	Parts 30^{o}	Parts 30^{o}	Parts 30^{o}
F-ratio 1 (MS_{With}/MS_{Bet})	0.37	0.34	0.41	0.34	0.30
Probability for F-ratio 1	1.00	1.00	1.00	1.00	1.00
F-ratio 2 (MS_{Bet}/ MS_{With})	2.73	2.94	2.42	2.93	3.34
Probability for F-ratio 2	<.001	<.001	<.001	<.001	<.001
Inference, Class level	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal
School District level					
Eta correlation between	0.263	0.237	0.197	0.320	0.321
Eta correlation within	0.965	0.971	0.980	0.947	0.947
E-ratio	0.273	0.244	0.201	0.338	0.339
E-ratio Inference	Parts 30^{o}	Parts 30^{o}	Parts 30^{o}	Parts 30^{o}	Parts 30^{o}
F-ratio 1 (MS_{With}/MS_{Bet})	0.38	0.44	0.72	0.24	0.23
Probability for F-ratio 1	1.00	1.00	0.95	1.00	1.00
F-ratio 2 (MS_{Bet}/ MS_{With})	2.62	2.28	1.38	4.20	4.41
Probability for F-ratio 2	<.001	<.001	.053	<.001	<.001
Inference, District level	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal
Overall MLA Inference	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal

Chapter 4

Background, Methods and Results for Research Questions 2 and 3

Research Question 2: To what extent do teacher and classroom characteristics affect the multilevel factor structure for each EDI scale?

Research Question 3: To what extent do teacher and classroom characteristics moderate the relative amounts of within-class and between-class variance for each EDI scale?

4.1 Background – Applying Frameworks to the EDI

The examination of teacher and classroom effects as part of the process of muiltilevel construct validation is a good example of what Zumbo (2007b) describes as third-generation differential item functioning (DIF). The first two generations of DIF were concerned with the identification of effects on items, and then the development of methodologies for DIF detection. In the third and current generation of DIF, the traditional spotlight on item structure is widening to also include elements of the testing context (e.g., teacher characteristics, number of students in the class) that may contribute to DIF. This brings the examination of DIF into the realm of Zumbo's (2007a) concept and purpose of validation – as one of the sources of inferential evidence needed to explain item response patterns.

Potential teacher bias in EDI ratings is challenging to measure because typically only one teacher rates each child. Interrater reliability studies for the EDI (Janus et al., 2000; Janus & Offord, 2007) compared teacher ratings with those of parents or early childhood educators, rather than with other teachers. Beswick, Sloat, & Willms (2003) addressed this challenge by defining biased ratings using the difference between EDI ratings by kindergarten teachers and

standardized assessment scores. However, it is not clear that this difference score was psychometrically sound, given its hybrid nature.

There has been no research until now that relates teacher and classroom characteristics to potential biases in EDI scoring, despite the obvious potential for systematic construct-irrelevant variation that comes with using teachers as raters. The smallest HELP neighbourhoods are theoretically most at-risk, as one teacher's biased ratings would have a disproportionately large effect on aggregated EDI scores. One potentially important source of construct-irrelevant variance relates to how well teachers know their students at the time when ratings are made. Less knowledge may translate into more relative scoring, as was shown earlier for Cycle 1 EDI data.

4.2 Methods– Applying Frameworks to the EDI

4.2.1 Research Question 2

This research question extends the multilevel factor structure analyses conducted in Chapter 3. The analytic strategy employed is analogous to that of hierarchical multiple regression, where blocks of theoretically relevant variables are added to investigate if and how the additions affect the model. In this case, the six classroom and teacher variables all exist at the classroom level. Therefore, what is being tested is their effect on the part of the model concerned with between-classroom variation. In keeping with the tradition of testing hierarchical models, an omnibus test for the block of six covariates is conducted first, to establish whether the covariates as a whole improve the model significantly. This is analogous to testing the statistical significance of the R^2 change in hierarchical multiple regression. Further interpretation of the influence of individual covariates is limited to situations where the omnibus test is statistically significant.

As before, due to the binary or three-category structure of the 103 core EDI items, the appropriate estimation technique for the two-level CFA is maximum likelihood (ML). The same sandwich estimation type of ML was employed to provide robust standard error estimates under the extant conditions of non-normality and nonindependence. The output from a two-level CFA using this ML estimator does not include the usual wide range of fit indices, only log-likelihood (including the scaling correction factor) and Information Criteria statistics.

As discussed in Chapter 3, absolute model fit cannot be assessed using the available log-likelihood and Information Criteria statistics. However, relative model fit can be assessed by comparing the "conditional" multilevel model (i.e., with classroom-level covariates) with the corresponding "unconditional" multilevel model (i.e., without classroom-level covariates). A chi-square difference test for this purpose can be computed using the log-likelihood values and scaling correction factors for the two nested models, as described on the Mplus website (www.statmodel.com/chidiff.shtml). Diagrammatic representations of generic unconditional and conditional two-level models are shown in Figures 1 and 2 below. They illustrate that the models differ only in the inclusion of classroom-level covariates that can potentially affect only the between-classroom latent variable. The "real" models used in the results that follow are EDI scale-specific, and so the number of items differs with each scale. However, in each conditional model, the same block of six covariates was used.

Figure 1. Unconditional Two-Level CFA Model

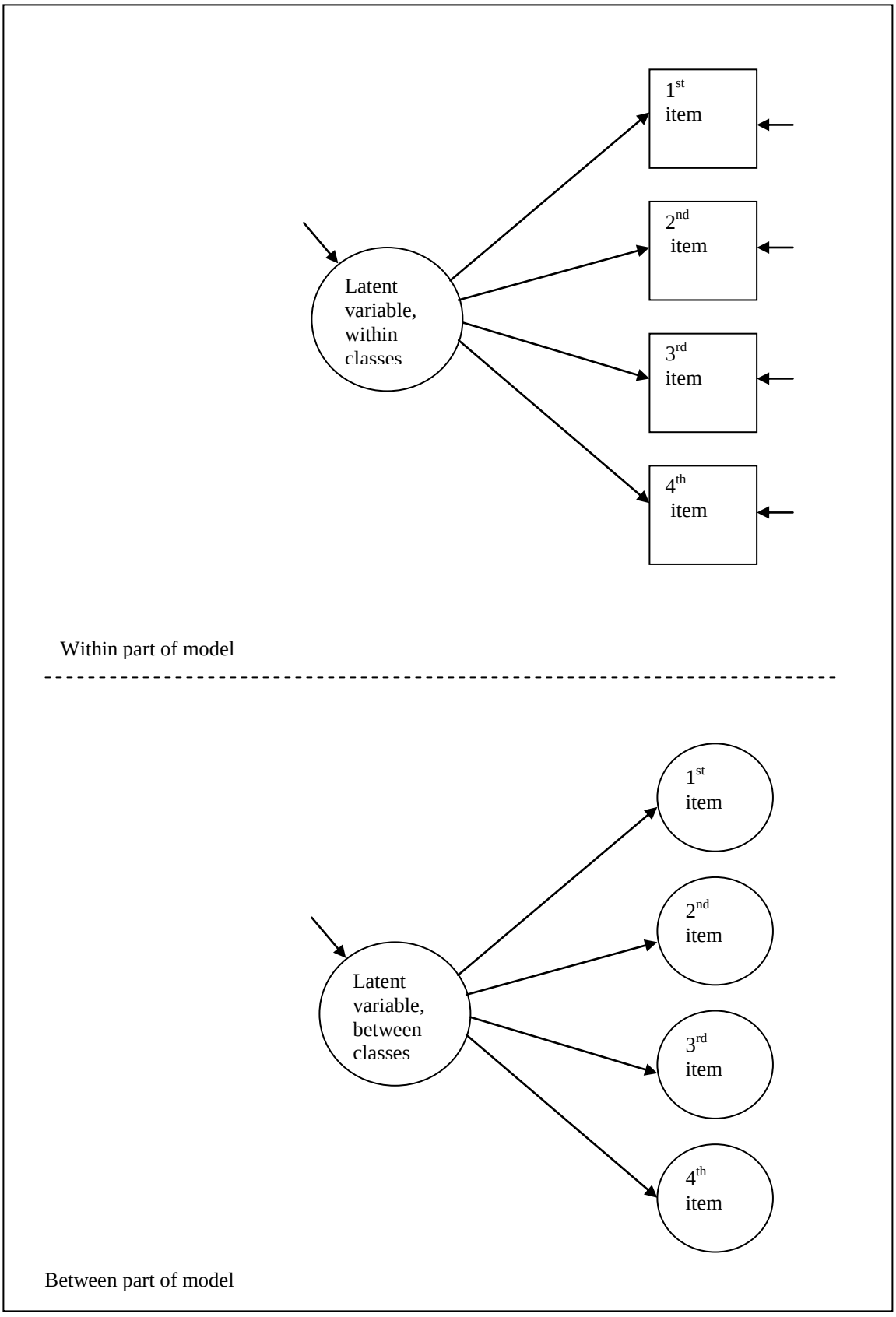
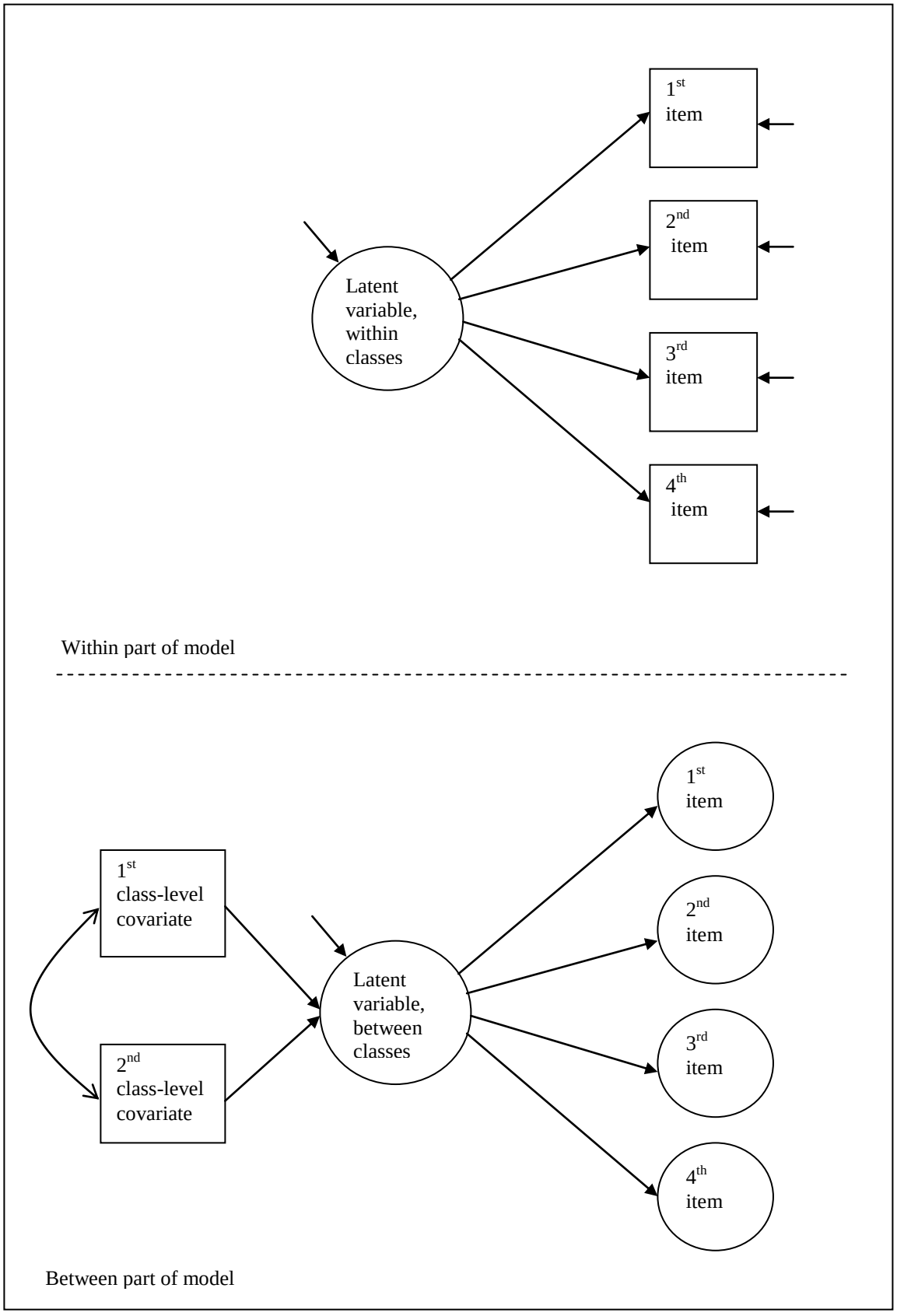


Figure 2. Conditional Two-Level CFA Model



In all of the model results presented below for each EDI scale, item and covariate coefficients as well as model variances are reported without their associated standard errors and *z*-scores. Instead, a symbol (*) is used to indicate statistical significance at the .05 probability level for these parameters. All of the item coefficients in the within-classroom portions of the models are statistically different from zero (with the obvious exception of the items fixed to 1), as would be expected with the large sample sizes involved. For the item coefficients in the between-level portion of the model, only a few (mostly in the language and cognitive development scale) are not statistically significant.

For teacher education, which was recoded to a three-category ordinal variable, it was necessary to create dummy variables for the purposes of the multilevel analyses below. (Teacher gender was already dummy-coded.) Two dummy variables were created, with *Bachelor degree but no graduate training* as the omitted category. Therefore, one dummy variable compared those teachers with no Bachelor degree with those in the omitted category, and the other dummy variable compared those teachers with some graduate work or a graduate degree with those in the omitted category.

4.2.2 Research Question 3

The same teacher and classroom characteristics can also potentially play a moderating role in whether the appropriate multilevel inference (in WABA parlance) is one of "wholes", "parts", or "equivocal." In Chapter 3, two different WABA analyses were conducted to investigate the appropriate inferences when considering only the individual and classroom levels (Single Level Analysis) as well as across the multiple levels of individuals, classes and school districts (Multiple Level Analysis). In summary, both of these analyses indicated that EDI inferences are most appropriate at the individual student level. The following research question

goes one step farther by asking the functional question about whether these WABA inferences have any boundary conditions, as defined by various classroom and teacher characteristics. Essentially, these models are testing for interactions between classroom/teacher characteristics and the inferences made for the whole sample.

4.2.2.1 Teacher and Classroom Moderators Tested

Five classroom and teacher characteristics were examined using WABA Multiple Relationship Analysis (MRA) for evidence that they may moderate the relative proportion of within-class and between-class variance in EDI scale scores. The five characteristics were: number of children in the classroom, teacher gender, teacher age, teacher experience with kindergarten students, and the educational attainment of the teacher. In keeping with MRA analysis, where moderators are conceptualized as boundary conditions, the variables that were measured on a continuous scale (i.e., number of children in the class and months of experience as a kindergarten teacher) were converted to categorical variables, each with four values. For both of these variables, the categories were chosen so that there were an approximately equal number of teachers in each category. The four categories for number of children in the class were: 10 or fewer, 11 to 16, 17 to 19, and 20 children or more. The four categories for months of experience as a kindergarten teacher were: less than 30, 30 to 70, 71 to 150, and 151 months or more.

4.2.2.2 Use of the E-ratio in MRA

In MRA, there are two main possible inferences that can be made about the pattern of variability across categories of a group-level moderating variable. If the same induction of wholes or parts (based on the E-ratio) applies regardless of the category of the moderator, then that induction is labeled by Dansereau et al. (1984) as "multiplexed," reflecting a lack of moderation. However, if a wholes or parts inference is made for one category of the moderator, and an equivocal inference is made for another category, then the induction is labeled as "contingent" on the value of the moderator.

4.3 Results

4.3.1 Research Question 2

4.3.1.1 Physical Health and Well-Being

Table 34 shows the results for both the unconditional and conditional multilevel CFA models. One key finding for this scale is that the amount of variance in the between-classroom factor for the unconditional model is not statistically significant from zero ($z = 1.39$, $p = .16$). This means that there was very little classroom-level variation available for modeling by the six covariates. According to the chi-square difference test, the effect of adding the six covariates was statistically significant, $\chi^2(6, N = 21{,}504) = 47.9$, $p < .0001$, with only teacher age reaching significance, $z = -2.61$, $p = .009$. Higher physical health and well-being scores were associated with younger teachers. However, the relationship between physical health and well-being scores and teacher age is not particularly meaningful because it accounts for an overall amount of between-classroom variance itself is not significantly larger than zero.

Table 34 Two-level CFA With and Without Teacher/Classroom Covariates, Physical Health and Well-Being Scale

	Coefficients - Unconditional Model		Coefficients – Conditional Model	
Item	*Within*	*Between*	*Within*	*Between*
A2	1.000	1.000	1.000	1.000
A3	1.629*	2.162*	1.626*	2.204*
A4	0.659*	0.346	0.659*	0.333
A5	1.322*	2.222*	1.322*	2.238*
A6	2.773*	3.600*	2.759*	3.721
A7	2.699*	3.673*	2.683*	3.803*
A8	3.837*	5.557*	3.823*	5.713*
A9	5.564*	10.446*	5.514*	10.791*
A10	9.036*	23.963*	8.952*	24.681*
A11	4.527*	21.922*	4.505*	22.618*
A12	2.847*	13.588*	2.835*	14.006*
A13	5.215*	21.686*	5.193*	22.382*
C58	1.051*	1.479*	1.045*	1.533*
Covariate				
Number in class				0.002
Gender				0.010
Age				-0.023*
Experience				0.004
Educ. – No Bach.				0.006
Educ. – Grad.				0.018
Factor Variance	0.400*	0.019	0.404*	0.018

* $p < .05$

4.3.1.2 Social Competence

In contrast to the results for the physical health and well-being scale, the amount of between-classroom variance in the unconditional model for social competence (see Table 35) was significantly greater than 0 ($z = 2.25$, $p = .02$). Thus, the classroom-level covariates have available some non-zero variance that they can potentially help to model. However, there was no statistically significant improvement in model fit by adding the block of six covariates, $\chi^2(6, N = 24{,}163) = 8.4$, $p = .21$. Therefore, the influence of each individual covariate was not interpreted.

Table 35 Two-level CFA With and Without Teacher/Classroom Covariates, Social Competence Scale

	Coefficients - Unconditional Model		Coefficients – Conditional Model	
Item	*Within*	*Between*	*Within*	*Between*
C1	1.000	1.000	1.000	1.000
C2	1.084*	1.051*	1.084*	1.042*
C3	1.367*	1.481*	1.368*	1.474*
C4	0.915*	1.433*	0.915*	1.429*
C5	1.856*	1.427*	1.857*	1.416*
C6	1.408*	1.360*	1.409*	1.351*
C7	1.307*	0.848*	1.308*	0.840*
C8	0.551*	1.015*	0.551*	1.011*
C9	1.225*	1.081*	1.226*	1.075*
C10	1.269*	1.303*	1.270*	1.295*
C11	1.316*	1.246*	1.317*	1.237*
C12	1.146*	0.812*	1.147*	0.804*
C13	1.488*	1.231*	1.489*	1.220*
C14	0.785*	0.840*	0.785*	0.833*
C15	0.928*	0.963*	0.929*	0.956*
C16	1.236*	1.382*	1.237*	1.372*
C17	0.701*	0.926*	0.701*	0.920*
C18	0.629*	2.973*	0.629*	2.964*
C19	0.446*	3.754*	0.446*	3.745*
C20	0.531*	3.828*	0.531*	3.819*
C21	0.601*	3.410*	0.602*	3.403*
C22	0.867*	1.230*	0.867*	1.222*
C23	0.957*	1.361*	0.957*	1.353*
C24	1.229*	0.946*	1.230*	0.938*
C25	1.015*	1.282*	1.015*	1.275*
C27	0.747*	1.127*	0.747*	1.121*
Covariate				
Number in class				-0.001
Gender				-0.086
Age				-0.029
Experience				0.028*
Educ. – No Bach.				-0.012
Educ. – Grad.				0.050
Factor Variance	6.666*	0.312*	6.665*	0.312*

* $p < .05$

4.3.1.3 Emotional Maturity

For the emotional maturity scale, the amount of between-classroom variance in the unconditional model was significantly greater than zero ($z = 11.13$, $p < .0001$). The effect of adding the six covariates was also statistically significant, $\chi^2(6, N = 18{,}185) = 13.2$, $p = .0407$. As Table 36 shows, teachers with graduate training or a graduate degree were associated with higher classroom emotional maturity scores than those with Bachelor degrees and no graduate work, $z = 3.11$, $p = .0019$.

Table 36 Two-level CFA With and Without Teacher/Classroom Covariates, Emotional Maturity Scale

	Coefficients - Unconditional Model		Coefficients – Conditional Model	
Item	*Within*	*Between*	*Within*	*Between*
C28	1.000	1.000	1.000	1.000
C29	1.067*	0.823*	1.067*	0.822*
C30	0.975*	1.113*	0.975*	1.113*
C31	1.171*	1.074*	1.171*	1.074*
C32	1.103*	1.280*	1.103*	1.280*
C33	1.052*	0.896*	1.052*	0.896*
C34	0.845*	0.842*	0.845*	0.841*
C35	1.117*	1.446*	1.117*	1.446*
C36	0.216*	-0.107*	0.216*	-0.108*
C37	0.936*	-0.221*	0.936*	-0.222*
C38	0.854*	-0.151*	0.853*	-0.152*
C39	1.058*	-0.161*	1.057*	-0.162*
C40	0.933*	-0.151*	0.933*	-0.152*
C41	0.773*	-0.111*	0.773*	-0.111*
C42	1.164*	-0.199*	1.164*	-0.200*
C43	1.253*	-0.182*	1.253*	-0.184*
C44	1.109*	-0.210*	1.108*	-0.212*
C45	1.308*	-0.129*	1.308*	-0.130*
C46	0.894*	-0.086*	0.894*	-0.087*
C47	1.144*	-0.184*	1.143*	-0.185*
C48	1.067*	-0.199*	1.066*	-0.200*
C49	1.355*	-0.144*	1.354*	-0.146*
C50	1.128*	-0.151*	1.128*	-0.152*
C51	0.510*	-0.121*	0.510*	-0.122*
C52	0.359*	-0.155*	0.359*	-0.155*
C53	0.298*	-0.173*	0.298*	-0.174*
C54	0.444*	-0.110*	0.444*	-0.111*
C55	0.476*	-0.168*	0.475*	-0.169*
C56	0.606*	-0.091*	0.606*	-0.092*
C57	0.103*	-0.047*	0.103*	-0.047*
Covariate				
Number in class				-0.009
Gender				-0.135
Age				0.017
Experience				0.006
Educ. – No Bach.				-0.093
Educ. – Grad.				0.368*
Factor Variance	5.687*	2.416*	5.692*	2.384*

* $p < .05$

4.3.1.4 Language and Cognitive Development

Similarly to the physical health and well-being scale, the amount of between-classroom variance for the language and cognitive development scale was not statistically distinguishable from zero ($z = 1.65$, $p = .10$). Therefore, any significant covariate effects found would still be relatively unimportant, given the insignificant amount of between-classroom variation.

The effect of adding the six covariates was statistically significant, $\chi^2(6, N = 20{,}472) = 25.5$, $p = .0003$. Of the six covariates, the two teacher education variables reached significance, showing a non-linear pattern (see Table 37). Not having a Bachelor degree was associated with higher language and cognitive development classroom scores compared to having a Bachelor degree with no graduate training, $z = 2.34$, $p = .02$. Having done graduate work was also associated with higher scale scores compared to having a Bachelor degree and no graduate training, $z = 2.20$, $p = .03$. However, these relationships are not particularly meaningful because they account for an insignificant amount of between-classroom variance.

Table 37 Two-level CFA With and Without Teacher/Classroom Covariates, Language and Cognitive Scale

	Coefficients - Unconditional Model		Coefficients – Conditional Model	
Item	*Within*	*Between*	*Within*	*Between*
B8	1.000	1.000	1.000	1.000
B9	0.702*	0.942*	0.698*	0.808*
B10	0.858*	1.473*	0.851*	1.241*
B11	1.179*	1.125*	1.170*	0.958*
B12	1.143*	0.935*	1.138*	0.786*
B13	0.879*	0.717*	0.875*	0.614*
B14	0.906*	2.318*	0.900*	1.951*
B15	1.106*	4.166*	1.100*	3.498*
B16	0.893*	5.227*	0.888*	4.392*
B17	0.916*	5.649*	0.912*	4.735*
B18	0.501*	2.904*	0.498*	2.449*
B19	0.846*	1.817*	0.843*	1.536*
B20	0.529*	1.909*	0.527*	1.608*
B21	0.942*	1.795*	0.933*	1.531*
B22	0.961*	10.837*	0.955*	9.125*
B23	0.867*	14.847*	0.864*	12.508*
B24	0.778*	0.835*	0.773*	0.711*
B25	0.885*	1.204*	0.879*	1.028*
B26	0.797*	0.907*	0.792*	0.778*
B27	1.064*	0.778*	1.056*	0.680*
B28	1.208*	0.826	1.200*	0.705
B29	1.028*	0.433	1.023*	0.372
B30	1.088*	0.253	1.080*	0.225
B31	1.105*	0.321	1.100*	0.275
B32	0.852*	0.298	0.844*	0.270
B33	0.772*	0.966*	0.768*	0.825*
Covariate				
Number in class				0.001
Gender				-0.058
Age				-0.012
Experience				0.013
Educ. – No Bach.				0.108*
Educ. – Grad.				1.159*
Factor Variance	9.021*	0.075	9.146*	0.104

* $p < .05$

4.3.1.5 Communication and General Knowledge

For the communication and general knowledge scale, the amount of between-classroom variance in the unconditional model was significantly greater than zero ($z = 11.19$, $p < .0001$). Adding the six covariates significantly improved the model, $\chi^2(6, N = 24{,}774) = 79.6$, $p < .0001$. As Table 38 shows, both increasing number of kindergarten children in the classroom, $z = 3.55$, $p = .0004$, and decreasing teacher age, $z = -6.35$, $p < .0001$, were associated with higher communication and general knowledge classroom scores.

Table 38 Two-level CFA With and Without Teacher/Classroom Covariates, Communication and General Knowledge Scale

	Coefficients - Unconditional Model		Coefficients – Conditional Model	
Item	*Within*	*Between*	*Within*	*Between*
B1	1.000	1.000	1.000	1.000
B2	0.549*	0.792*	0.547*	0.788*
B3	0.866*	0.894*	0.864*	0.895*
B4	0.361*	0.628*	0.361*	0.627*
B5	0.807*	0.902*	0.806*	0.897*
B6	0.673*	0.859*	0.672*	0.859*
B7	0.484*	0.485*	0.484*	0.484*
C26	0.375*	0.262*	0.375*	0.262*
Covariate				
Number in class				0.050*
Gender				-0.409
Age				-0.619*
Experience				0.028
Educ. – No Bach.				0.515
Educ. – Grad.				0.315
Factor Variance	32.206*	9.904*	32.206*	9.177*

* p < .05

4.3.2 Research Question 3

Tables 39 to 43 show, for each of the five teacher/classroom variables separately, the E-ratios for each category across the five EDI scales. For all five EDI scales, an induction of a parts situation was found for all categories in teacher gender, teacher age, experience as a kindergarten teacher, and teacher education. Thus, a parts induction applies across conditions for all of these variables, implying that the "parts" inference is multiplexed. The number of kindergarten children in the classroom (Table 39), however, did show a pattern of contingent induction for all five EDI scales. Specifically, an inference of an equivocal situation is made when the class size was 10 or less, while the usual inference of parts applied to the other categories of this variable.

Table 39 E-Ratios for Number of Students in the Class, by EDI scale

	E-ratio				
	Phy	Soc	Emo	Lan	Com
10 or fewer	.785*	.770*	.778*	.824*	.870*
11 to 16	.640+	.569+	.623+	.568+	.643+
17 to 19	.598+	.506+	.575+	.583+	.643+
20 or more	.616+	.543+	.559+	.625+	.640+

Induction: * Equivocal + Parts

Table 40 E-Ratios for Teacher Gender, by EDI Scale

	E-ratio				
	Phy	Soc	Emo	Lan	Com
Female	.634+	.562+	.595+	.612+	.663+
Male	.596+	.587+	.616+	.569+	.529+

Induction: * Equivocal + Parts

Table 41 E-Ratios for Age Group of Teacher, by EDI Scale

	E-ratio				
	Phy	Soc	Emo	Lan	Com
20 to 29	.627+	.578+	.620+	.654+	.670+
30 to 39	.661+	.548+	.552+	.606+	.670+
40 to 49	.603+	.561+	.611+	.585+	.633+
50 or older	.640+	.569+	.603+	.626+	.655+

Induction: * Equivocal + Parts

Table 42 E-Ratios for Experience as a Kindergarten Teacher in Months, by EDI Scale

	E-ratio				
	Phy	Soc	Emo	Lan	Com
Less than 30	.649+	.558+	.555+	.654+	.610+
30 to 70	.664+	.560+	.597+	.611+	.704+
71 to 150	.582+	.556+	.601+	.565+	.652+
Over 150	.634+	.575+	.632+	.613+	.660+

Induction: * Equivocal + Parts

Table 43 E-Ratios for Educational Attainment of Teacher, by EDI Scale

	E-ratio				
	Phy	Soc	Emo	Lan	Com
No Bachelor degree	.711+	.696+	.666+	.671+	.702+
Bachelor degree, but no graduate school	.623+	.550+	.586+	.599+	.654+
Some graduate school or graduate degree	.654+	.561+	.606+	.648+	.667+

Induction: * Equivocal + Parts

To look more closely at this moderating effect of class size, each E-ratio was decomposed into its numerator and denominator, so that between and within variance could be compared across the four moderator categories. Table 44 shows this variance breakdown for the four categories, for each of the EDI scales. A similar pattern can be seen regardless of the EDI scale. The within-group variance is relatively consistent across categories of the moderator, while there is a greater range of variances between groups. Of particular interest for the between-group variances, the variance for the first category (10 students or less) is higher than for the other three categories.

Table 45 shows these two different patterns for the between and within variances using the F_{max} statistic, which is the ratio of the highest variance to the lowest variance across categories. For the between-group variances, F_{max} ranges from 1.63 for the communication and general knowledge scale to 2.24 for the social competence scale. These F_{max} statistics, with degrees of freedom ranging from 565 to 703 for both numerator and denominator, are all

strongly statistically significant (all *p*s < .0001). This confirms the inference that the between-group variance across categories is not homogeneous.

As for the within-group variances, the F_{max} statistics are visibly smaller, ranging from 1.06 to 1.14. Unfortunately, statistical inferences using these within-group F_{max} statistics cannot be reasonably made due to the extremely large degrees of freedom - greater than 10,000 for all numerators and denominators. Under these circumstances, any F_{max} greater than 1 would be statistically significant. The observed within-group F_{max} statistics are too small in absolute value to be considered practically significant. Looking back at the actual pattern of variances shown in Table 44, within-group variance is relatively consistent across categories, compared to the greater heterogeneity in the between-group variance.

Therefore, the above pattern of results suggests that the only meaningful variance differences are at the between-group level. These between-group differences are explored in Table 46, which shows all six pairwise comparisons of the four categories of class size, for each of the EDI scales. Each comparison involved taking the ratio of the higher variance to the lower variance, which can be tested statistically using an F-ratio. Again, the numerator and denominator degrees of freedom range from 565 to 703. To protect the family-wise error rate for these pairwise analyses, a Bonferroni correction was applied, adjusting the value of alpha to .0083 (.05 divided by 6). As Table 46 shows, all pairwise comparisons involving the lowest category (10 kindergarten students or less) were highly significant (F-ratios ranging from 1.60 to 2.24). All other pairwise comparisons had distinctly lower F-ratios, ranging from 1.00 to 1.21; all of these failed to reach statistically significance using the adjusted critical alpha.

Taken all together, this pattern of results (which applies to all five EDI scales) can be summarized as follows. The WABA MRA analysis indicates that only one of the teacher/classroom variables moderates the pattern of variation enough to change the appropriate

inference. That variable is the number of kindergarten children in the class. When the number of children is small (10 or fewer), the appropriate inference is “equivocal,” meaning that within-group and between-group variation are somewhat balanced. When the number of children is higher than 10, the appropriate inference is “parts,” which means that there is a preponderance of within-group variation. Looking more closely at the variation patterns across categories, there actually is not much difference in the amount of within-group variation. There is, however, significantly greater between-group variation when the number of kindergarten children per class is relatively small. Therefore, teachers with 10 or fewer students were more likely than other teachers to rate students absolutely rather than relatively.

Table 44 Variances Between and Within for Each Category of Number of Students in the Class, by EDI Scale

Number of students in the class	Variance									
	Physical		***Social***		***Emotional***		***Language***		***Communication***	
	Between	*Within*	*Between*	*Within*	*Between*	*Within*	*Between*	*Within*	*Between*	*Within*
10 or fewer	0.912	1.479	1.641	2.764	1.033	1.704	1.868	2.754	3.491	4.616
11 to 16	0.624	1.522	0.891	2.755	0.687	1.770	0.970	3.001	2.147	5.187
17 to 19	0.527	1.471	0.734	2.868	0.613	1.854	0.956	2.813	2.181	5.276
20 or more	0.545	1.436	0.859	2.912	0.590	1.889	1.108	2.835	2.152	5.250

Table 45 F_{max} (between and within) for Number of Students in the Class, by EDI Scale

	Fmax				
	Phy	***Soc***	***Emo***	***Lan***	***Com***
F_{max} between	1.73	2.24	1.75	1.95	1.63
F_{max} within	1.06	1.06	1.11	1.09	1.14

Table 46 Pairwise Comparison F Tests (Between Variation Only) for Number of Students in the Class, by EDI Scale

	E-ratio									
	Physical		***Social***		***Emotional***		***Language***		***Communication***	
	F-ratio	*Prob.*	*F-ratio*	*Prob.*	*F-ratio*	*Prob.*	*F-ratio*	*Prob.*	*F-ratio*	*Prob.*
10 or fewer vs. 11 to 16	1.46	<.001	1.84	<.001	1.50	<.001	1.93	<.001	1.63	<.001
10 or fewer vs. 17 to 19	1.73	<.001	2.24	<.001	1.69	<.001	1.95	<.001	1.60	<.001
10 or fewer vs. 20 or more	1.67	<.001	1.91	<.001	1.75	<.001	1.69	<.001	1.62	<.001
11 to 16 vs. 17 to 19	1.18	.020	1.21	.009	1.12	.079	1.01	.452	1.02	.403
11 to 16 vs. 20 or more	1.14	.050	1.04	.312	1.16	.032	1.14	.050	1.00	.490
17 to 19 vs. 20 or more	1.03	.362	1.17	.030	1.04	.320	1.16	.038	1.01	.453

Chapter 5

Discussion

5.1 Review of Motivation and Purpose

As is the case for many methodological studies, this dissertation research has been motivated and shaped by the interplay of general and specific interests. For the past 20 years, I have been a statistical consultant specializing in early child development (ECD) research, as well as an occasional post-secondary instructor of research methods courses. Over this time, I have developed a general interest in the validity and reliability of ECD-related measures. My first exposure to the Early Development Instrument dates back to 1999, when I was contracted to analyze the results of the first EDI pilot study in British Columbia, for children in the Vancouver School District. Five years later, as I was modelling EDI-SES associations for the *British Columbia Atlas of Child Development* (Kershaw et al., 2005), I read the limited (and unpublished) psychometric literature for the EDI, and realized that none of it took the multilevel nature of the school readiness construct into account. Given that EDI implementation was expanding quickly across Canada and internationally, it became clear that validating the EDI as a multilevel construct was both timely and necessary. This realization has motivated me to explore current methodological approaches to multilevel construct validation that could be meaningfully applied to the EDI.

Accordingly, the purpose of this dissertation research was to articulate two different methodological frameworks for conducting multilevel construct validation, and then apply them in the context of the EDI. These two frameworks originally came from the field of organizational psychology, in which validity issues relating to multilevel constructs and their appropriate level of interpretation have been addressed in their literature for at least 25 years. Relatively little attention has been paid in the educational literature to validity issues relating to the use of

multilevel constructs. Educational effectiveness studies, for example, typically do not provide any overt rationale for the level(s) of aggregation at which their measures are analyzed (Griffith, 2002). Therefore, this dissertation is an effort to translate, adapt, and expand upon the knowledge about multilevel construct validation techniques from organizational psychology to one of the realms of education, school readiness, using the increasingly-popular EDI measure as a case in point.

This translation is not simply the rote exercise of following a series of predetermined analyses, requiring only the substitution of school-related terms for their workplace analogues, such as "class" for "work group," or "social competence" for "people skills." Any validation process requires careful consideration of the theory behind a particular measure, its psychometric characteristics, the desired levels of aggregation, and so on (Zumbo, 2007a, 2009). Thus, these techniques must be applied mindfully even for studies within the same academic discipline. When translating validation techniques across disciplinary borders, as is the case for this dissertation, some unique elements will exist, requiring even more careful consideration. For example, in the realm of developmental psychology in general and school readiness in particular, measures of risk or vulnerability are particularly salient, as is certainly true for the EDI. Therefore, it is important to identify and address any resulting specific methodological challenges for the multilevel validation of vulnerability measures.

5.2 Novel Contributions

There are two levels of novel contributions of this dissertation. The primary contribution is to the broad methodological inquiry that articulates two techniques of multilevel construct validation and their adaptation to school readiness measures. This includes establishing the importance of a multilevel focus in the realm of school readiness, adapting methodological tools

to take measurement issues into account, and demonstrating how the analytic tools used for multilevel construct validation can be used in this realm. The secondary contribution of this dissertation is the knowledge gained about the psychometric properties of the EDI as a result of adapting and using these tools for multilevel construct validation.

5.3 Two Approaches to Multilevel Validation

Both of the analytical approaches explored in this dissertation (i.e., WABA and mixed models via HLM or SEM) share the same general goal of determining when aggregation of individual scores is justified, based on patterns of within-group and between-group variability. Either homogeneity of individuals within groups or heterogeneity between groups (or both) is taken as evidence, by both approaches, that aggregation is justified (Klein et al., 2000). Beyond this commonality, the two approaches have quite different foundational assumptions. The most important is that WABA allows for a third multilevel pattern – interdependent individuals within groups. WABA practitioners also emphasize practical significance, arguing for their E-ratio because it is unaffected by sample size (Dansereau & Yammarino, 2004, 2006). In response, the advocates of ICC(1) and ICC(2) argue that eta-squared (the building block of the E-ratio) is sensitive to group size (Bliese & Halverson, 1998; Chen et al., 2004b). These issues about the influence of group size and number of groups on these statistics are difficult to resolve, as they are based on different assumptions about the appropriateness of modeling within-group variation (Hofmann & Jones, 2004).

Within this healthy debate about the merits of each approach to multilevel construct validation (e.g., Chen et al., 2004b; Kim, 2004), there is also recognition that these approaches can be complementary (Dansereau & Yammarino, 2000; Klein et al., 2000). WABA is unique in its general purpose of assessing the appropriate level(s) of aggregation for variables, rather than

assuming the levels beforehand. In contrast, HLM (Bryk & Raudenbush, 1992) and SEM models assume that variables expressed at particular levels of aggregation are valid. WABA can therefore be used as a screening technique to establish, for the variables in the model, the appropriate level(s) of aggregation for later HLM or SEM analyses. WABA does not attempt to assess causal claims (as in SEM) or predict variability in slopes (as in HLM); if these analyses were desired, they would follow WABA analyses using levels established as appropriate.

5.4 Main Validation Results for the EDI

This sometimes subtle complementarity of approach and purpose has been demonstrated in the WABA and SEM analyses conducted in this paper. The WABA analyses considered different levels of aggregation *a priori,* and concluded that individual scores are the appropriate level of analysis for each of the EDI scales. The multilevel SEM models assumed clustering at the class and neighbourhood levels, and then the effects of these clusters on model fit were analyzed. Model fit was not noticeably improved by clustering. Thus, these two approaches came to the same conclusion that a measurement model for EDI scale scores is best situated at the individual child level.

The strength and reliability of classroom-level effects were also assessed by calculating the intraclass correlation coefficients ICC(1) and ICC(2), respectively. As discussed in section 4.2, the ICC(1) values indicate that a substantial portion of the overall variance in the actual scores is attributable to the classroom level, and the ICC(2) values confirm that these means are reliable.

These two results (lack of a significant clustering effect, but a sizeable ICC(1)), may seem contradictory at first glance. The resolution lies in the different purposes of these two analyses. The SEM analyses are concerned with whether the items in an EDI scale come together

to represent one latent variable. By including a between-class variance component, the question is whether this unidimensional representation is improved by taking this clustering into account. The answer is that a unidimensional model fits very well for each scale whether or not between-classroom variance is included. This good-fitting unidimensional model justifies the calculation of a total or mean score, which in turn allows for the calculation of ICC(1) and ICC(2) statistics, which are based on the total score. The question that is answered by the ICC(1) statistic is how much of the total variance in the individual scores can be explained by class membership.

How consistent are these patterns of results with the EDI level of theory as described by its developers? According to Janus and Offord (2000, 2007), the EDI is theoretically situated in a social constructivist framework, but EDI scores should never be interpreted at the individual child level.

There are several key premises in a social constructivist philosophy of child development. One is that individual children are active, unique, and complex. Another is that process of learning and development is driven by the dynamic interplay over time of children with their many environments. Young children actively construct meaning by discovering new concepts, principles, and facts. Teachers can facilitate this process, both by challenging students with learning tasks slightly above their current developmental level (i.e., using Vygotsky's [1978] concept of the zone of proximal development), and by reinforcing the idea that teacher and student are mutually and equally involved in learning.

Therefore, one implication of situating the EDI within a social constructivist framework is that the theory must include the individual child level. Janus and Offord (2007, p. 5) are explicit that this is true for the EDI, which "is positioned in the context of a social constructivist approach by providing the 'child' component necessary to complete the whole picture of community-based school readiness." In terms of construct validation, this means that individual

variation is relevant to the construct, and so the construct must take individual differences into account. So far, this shows a good match between EDI level of theory and the results.

The apparent contradiction between embracing a social constructivist philosophy while disallowing individual-level interpretations of EDI scores needs to be understood in context. The philosophy reflects the growing dominance of interactionist models of development, such as Bronfenbrenner's (1977) ecological model, as well as the simultaneous availability of multilevel statistical techniques such as HLM (Bryk & Raudenbush, 1992) and structural equation modeling (SEM) for testing these models. The caution against individual-level interpretations reflects the population health orientation of the two major Canadian proponents of the EDI (the Offord Centre for Child Studies at McMaster University, and the Human Early Learning Partnership based at the University of British Columbia), but is also a reaction to the historically poor predictive validity of child-level school readiness measures (Ellwein et al., 1991; Shepard, 1997). Individual scores are meaningful, but philosophically and practically, the focus of potential score use and interpretation, as well as interventions, is at a community, rather than individual level.

Given this focus, it is surprising that until Janus and Offord's (2007) first attempt to conduct a multilevel validation study, all other EDI reliability and validity studies (see section 1.4) were conducted at the individual level only. The assumption seems to be that if validity and reliability are demonstrated at the individual level, then this would extend automatically to aggregated levels, regardless of what they are. This assumption allows the possibility of spurious inferences due to making an atomistic inferential error (Diez-Roux, 1998). More validation attention needs to be made to the validity of aggregated scores, if that is really the level of theory (i.e., inferences) for EDI scores.

A developmental construct like social competence certainly does not have the same meaning for individuals as it has for classes, schools, or neighbourhoods. For example, whatever genetic component of social competence accounts for individual differences between children, there can be no such component that distinguishes entities like schools. On the other hand, it is likely that the variables that relate to social competence at the individual level probably also relate similarly at higher levels of aggregation. This is what Bliese (2000) calls a "fuzzy composition model" for a multilevel construct, where the meaning is similar, but not identical, across levels. In other words, a completely new meaning for something like social competence does not emerge for schools or neighbourhoods.

5.5 Dimensionality of Aggregated EDI Scores

Finally, the results of the factor analyses at the class and neighbourhood levels show that two of the EDI scales (physical health and well-being, and emotional maturity) do not demonstrate essential unidimensionality at either aggregate level. For the physical health and well-being scale, a two-factor solution was preferred for both levels (with some items not loading on either factor), while for the emotional maturity scale, a three-factor solution made sense for both levels. This calls into question Janus and Offord's (2000, 2007) assertion that each EDI scale is unidimensionally meaningful, but only at a population (i.e., aggregated) level. The pattern of results in this study suggests, for these two EDI scales only, that subscale scores (based on the items in each factor) should be summarized at a population level, rather than the overall scale scores.

5.6 Teacher/Classroom Level Effects

How consistent is the finding of classroom-level influences in the range of 19% to 25% of total variance with the EDI developers' level of theory? On one hand, an implication of a social constructivist perspective is that the ideal level of theory is a multilevel one, implying influences on individual development from phenomena at a variety of levels of aggregation, such as the class, the school, the school district, and the neighbourhood. On the other hand, ICC(1) values of this range are perhaps unexpectedly high, as will be explained below.

The idea of using kindergarten teachers as raters makes good sense from a population health / educational policy perspective. Kindergarten is the first opportunity to capture data systematically and quickly from essentially all individual children in a jurisdiction. Teachers theoretically have had enough time to get to know each child well enough to provide accurate and informed scores on the 103 EDI core items. Population-level school readiness data, which can be linked to other large-scale data relating to children's development, makes possible the shift from the traditional individual differences approach to development, to the current emphasis on the social determinants of population health (Guhn, Janus, & Hertzman, 2007).

In addition to the utilitarian advantage, teachers' ratings of student outcomes, compared to direct measures by trained assessors, have demonstrated either advantages either in predictive validity, or in the ratio of predictive strength to required effort. In terms of predicting early academic achievement, Kim and Suen's (2003) meta-analysis showed that teacher ratings are superior to direct assessments. Duncan et al. (2007) concurred that teacher-rated school readiness has good predictive validity for school achievement, even seven years later. With respect to the EDI in particular, Forget-Dubois et al. (2007) found that EDI scores accounted for 36% of the variance in Grade 1 outcomes, compared to 50% for a battery of much more labour-intensive direct assessments.

Of course, comparing the EDI to a battery of standardized tests is a false dichotomy. There are other teacher-led assessment practices with demonstrated effectiveness, such as the Work Sampling System (WSS; Meisels, 1992). The WSS focuses on guiding students' learning, collecting direct evidence of student performance, and helping parents understand developmental progressions, rather than inferring students' abilities or making placement decisions. This intensive approach, however, means that the WSS is not designed to be useful from a population-health perspective.

For the EDI, the advantages of using teachers as raters are counterbalanced by two important trade-offs that potentially undermine its multilevel construct validity. First, consistent with a social constructivist point of view, teachers by definition are measuring the developmental status of their students including the influence that they as teachers have had on them. This influence, however short-lived compared to other influences, accounts for 19% to 25% of the total variation in children's scoring.

The second trade-off is that this between-class variation may be irrelevant to the construct, which would bias individual scores. Given the sizeable ICC(1) values after only six months of exposure to teacher influences, it seems likely that at least some of the teacher variation is construct-irrelevant. Examples of construct-irrelevant variation would be between-teacher variability in the leniency or severity of ratings, or a tendency to systematically rate particular subgroups of children using a different standard than other children. Unfortunately, distinguishing between construct-relevant and construct-irrelevant classroom-level variation is often not possible.

Two types of analyses were conducted to assess the potential effects (whether construct-relevant or irrelevant) of teacher and classroom characteristics for each EDI scale. The first analysis tested the improvement in fit of the two-level (individual and classroom) CFA models

after adding six teacher and classroom covariates. Assuming non-zero between-class variation available to be modeled, a chi-square difference test was used to test whether the teacher/classroom covariates significantly improved the between-class model fit. Where appropriate, this was followed up with *z* tests of individual covariates. The second analysis used the WABA approach to test for a moderation effect of any teacher or classroom variables on the most appropriate multilevel inference. As will be shown below, this type of analysis provides a unique opportunity to distinguish a construct-irrelevant effect.

5.7 Adding Teacher/Classroom Covariates to the Multilevel CFA

Of the five EDI scales, two (physical health and well-being, and language and cognitive development) had between-class variation that was not statistically distinguishable from zero, precluding any examination of effects at the classroom level. For the three EDI scales with non-zero between-class variation, significant teacher and/or classroom effects were found for the emotional maturity as well as the communication and general knowledge scales. Teachers with graduate training or a graduate degree gave higher emotional maturity scores than teachers with Bachelor degrees but no graduate school education. With regards to communication and general knowledge, younger teachers and those with more students in their classrooms gave higher scores than older teachers or those teaching fewer students. Although there was non-zero between-class variation for the social competence scale, the six classroom-level predictors together did not account for a significant amount of this variation.

There are very few studies examining teacher and classroom effects on development for children in this general age range, and only one study where multilevel modeling was used. In this study, Mashburn et al. (2006) found, for children in pre-kindergarten classes, levels of non-independence in the same range (23% to 26%) as found in Janus and Offord (2007) and in the

current study. Teacher ratings of social competence and problem behaviours (similar to emotional maturity for the EDI) were associated with a few teacher and classroom characteristics, after taking individual child characteristics into account. With regards to the specific teacher and classroom characteristics that were also considered in this study, higher social competence ratings were found for teachers with less teaching experience, and for classrooms with fewer students. Teachers with more teaching experience rated their students as having more behaviour problems. No teacher education effects were found. Teacher age and gender were not considered as potential predictors in their study.

In contrast to the Mashburn et al. (2006) results, the current study found that emotional maturity ratings were unrelated to teaching experience, but were higher for teachers with graduate-level education. More teaching experience, rather than less (as in the Mashburn study), was associated with higher scores on social competence, though this individual effect was not interpreted due to the lack of an omnibus covariate effect.

There are a number of potential explanations for the relationship between teacher education and emotional maturity. Perhaps more educated teachers are particularly adept, as a result of knowledge gained at graduate school, at providing learning opportunities for their students that relate to emotional development. Alternatively, it may be a selection effect – teachers with graduate training may be assigned preferentially to classes where students are particularly emotionally mature, or may choose to teach in neighbourhoods with higher socioeconomic status. However, one might expect that more experienced teachers would also have more capacity to choose their students, yet emotional maturity was not related to experience.

There were two teacher/classroom effects found for the communication and general knowledge scale, with younger teachers and larger class sizes associated with higher scores.

Younger teachers may have lower expectations for children on this domain than older teachers, who through life experience have come to appreciate the importance of good communication skills. As before, a selection effect may also come into play, with younger teachers more likely to work outside of the lower mainland of British Columbia, where there are far fewer ESL students. The positive association between class size and communication and general knowledge scale scores is more challenging to understand. Perhaps in smaller classes, where teachers have more opportunity to interact with students individually, teachers give more accurate ratings on this scale. In larger classes, teachers may give more "benefit of the doubt" to students whose communication skills they know less well.

The lack of consistency in results between the current study and Mashburn et al. (2006) is not surprising, given differences in how the outcome and explanatory variables were measured, the age of the children, and especially the types and purposes of models tested. For example, they controlled for individual child characteristics in an HLM model designed to test classroom-level predictors of social/emotional development, while in the present study had no individual-level covariates, and used two-level CFA to test whether the fit of the multilevel measurement model relates to classroom-level characteristics.

5.8 Teacher/Classroom Characteristics as Potential Moderators

No moderation effect was found for any of the teacher characteristics: age, gender, education or experience. However, the appropriate inference was contingent on class size. When class size was 10 or less, the E-ratio inference indicated "independent individuals," but when it was higher, this inference changed to "heterogeneous individuals." These results suggest that teachers with small class sizes were more likely to rate students absolutely rather than relatively.

This makes logical sense, as teachers should have a better opportunity to get to know their students when class size is small.

The findings in this study of more absolute teacher ratings when class sizes are small is consistent both with the educational literature on the benefits of smaller class sizes, and with a social constructivist view of effective teaching and learning. Absolute teacher ratings imply greater knowledge of students as individuals, which in turn should be positively related to the frequency and quality of teacher-student interactions. Indeed, many researchers (e.g., Blatchford, Bassett, Goldstein, & Martin, 2003; Kreiger, 2003; Pedder, 2006) have found that smaller classes are associated with more knowledge of pupils, and more one-to-one teaching and support. Consistent with an interactionist perspective, students are also more actively engaged in their learning when class sizes are small (Finn, Pannozzo, & Achilles, 2003) showing a greater likelihood of initiating and sustaining contact with teachers (Blatchford, Russell, Bassett, Brown, & Martin, 2007). Smaller classes provide a more intimate learning context where rules, relationships, and dynamics are qualitatively different than larger classes (Pellegrini & Blatchford, 2000).

Because the WABA approach allows for the possibility of groups defined as interdependent individuals, it also can suggest unique ways of understanding both school readiness scoring and potential strategies for improvement. Traditionally, intervention strategies have been perceived dichotomously – help can be targeted at individuals, and/or at groups. For example, the developers of the EDI have emphasized the latter, authorizing only group-level (e.g., neighbourhood or school district) interpretations and interventions based on EDI scores. This dichotomy is analogous statistically to what Dansereau and Yammarino (2004) call the "ANOVA perspective," where there are only two possible situations - independent individuals or groups with exchangeable individuals.

The third possible situation of interdependent individuals within groups, particularly in the context of formal education, pertains directly to the effect of teachers as raters of their students. Teachers' ratings of student characteristics relative to the class mean are clearly of theoretical interest in educational research (Snijders & Bosker, 1999). They lead directly to intervention strategies that go beyond targeting particular individuals or groups, as scoring depends on group composition. As was found in the present study, for class sizes higher than 10 students, teachers' EDI ratings become more relative in character. This situation calls for more global changes such as reducing class sizes and providing better or earlier EDI training to kindergarten teachers.

5.9 Relative Scoring is Construct-Irrelevant

Modern interaction-based conceptions of school readiness (e.g., Bronfenbrenner, 1979) would expect that a child's developmental status midway through kindergarten has been influenced through interactions with the teacher. From a holistic perspective, these interactions can relate in complex ways with various characteristics of both child and teacher. For example, a shy boy's social competence may experience sudden growth in kindergarten if his teacher helps him get involved in sports activities during recess. When this child is scored on the EDI item C4 ("is able to play with various children"), the fact that the teacher had a significant influence on that rating does not mean that score needs to be adjusted somehow (i.e., has some irrelevance built in). However, if the teacher rates that child relative to where the child was at the start of kindergarten, and rates other children in a more absolute way, then the score will have some construct irrelevance. Because teachers have had such integral and recent influence on children's developmental status, and they are also the raters on the EDI, it is difficult to distinguish between construct-relevant and construct-irrelevant sources of score variance. Thus, a finding that

younger teachers tend to give higher scores on the communication and general knowledge scale is not necessarily evidence of construct-irrelevant variance. It may be that younger teachers are, for example, more lenient in their ratings for this scale (irrelevant), or perhaps younger teachers tend to do a better job of encouraging the abilities pertaining to this scale (relevant).

Even within an interactionist framework of school readiness, the finding of more absolute scoring for small class sizes is more difficult to construe as construct-relevant variance. To be relevant, it would imply that children's developmental status would be more interdependent for children in larger classes, and more independent for those in small classes. If anything, from an interactionist perspective, the opposite trend would seem more likely. The construct-irrelevant explanation - that scoring is affected by how well teachers get to know their students – makes intuitive sense.

5.10 Methodological contribution – Development of a new fit index for multilevel SEM

In this dissertation, the development of a new fit index, which is given the name RMR-P (Root Mean Square Residual – Proportion) is described for the specification of multilevel SEM models with categorical data. The impetus for creating the RMR-P came from the untenable computational demands inherent in conducting multilevel covariance structure analyses for categorical data, particularly when there are more than a few items to be considered. Maximum likelihood techniques are needed to estimate the polychoric correlation matrices for these data, with one additional dimension of mathematical function that must be integrated for every additional item. These computational demands preclude calculating the pooled within-group or sigma-between covariance matrices that are required for the multilevel covariance structure analysis strategies proposed by Hox (2002) or Muthen (1994). Recently proposed solutions to overcoming these problems, such as the approach suggested by Grilli and Rampichini (2007), are

not appropriate for measures like the EDI with a relatively large number of items per scale, and item distributions far from Gaussian.

5.11 Potential Implications of the Findings

Before the modern era of unitary construct validity (e.g., Messick, 1989), the practical implications of validation studies were focused mostly on the test itself, usually in terms of its psychometric properties or criterion-related (i.e., concurrent or predictive) utility. Broader, more contextualized implications were often not considered to be part of the realm of validation. In contrast, the current unitary view focuses on theoretical explanations of test score variations, rather than descriptions. In such an explanatory system, it is also necessary to address additional implications like ethical, social, and policy consequences (Zumbo, 2009).

The inclusion of these consequences brings measurement work into the realm of social policy. The implications of good measurement are no longer the exclusive concern of measurement specialists, extending to a variety of stakeholders (Messick, 1995). In the school system, this would include teachers, principals, administrators, and government officials, all of whom have consequential roles to play in making inferences based on test scores, and designing appropriate interventions. For example, a principal may be relatively unconcerned about high rates of vulnerability on the EDI communications scale at her school, if the school has a high ESL population. With regards to the EDI, interest in how it measures school readiness has helped to establish early childhood education and early learning generally as important issues for those working in the formal school system.

5.11.1 General Cross-Level Implications

One of the implications of multilevel measurement, as Zumbo and Forer (in press) have noted, is that such tests can have consequences for policy decisions that can impact individual students, despite the intention to limit interpretation (and intervention) to the community level. One good example of this is the British Columbia government's stated goal of reducing the proportion of developmentally vulnerable children to under 15% by the year 2015 (Province of British Columbia, 2008), presumably based on EDI test results. This goal places the emphasis on individual-level developmental vulnerability, rather than the community-level vulnerability that is considered meaningful in the EDI view of school readiness. If instead, the provincial goal for 2015 had been for fewer than 15% of the 478 provincial neighbourhoods to have vulnerability rates above a particular threshold, this would be more consistent with the view of the EDI as community-level measure. It would also put the onus on community-level interventions. The provincial goal, as currently defined, inadvertently encourages interventions designed to impact individual students. By focusing on individual students rather than communities, there is also the possible unintended consequence that the goal could be achieved by focusing interventions or resources on more population-rich parts of the province, like the Greater Vancouver area.

Despite these undesirable cross-level consequences, there are some positive implications that flow from the establishment of the psychometric meaningfulness of individual-level EDI scores. First, it validates HELP's goal of understanding developmental trajectories from birth onwards, which is based on linking individual-level data (including the EDI) from a variety of population-based databases, such as those kept by the Ministries of Health, Education, and Children and Family Development. The EDI is already being linked with Foundation Skills Assessment (FSA) scores to create an index of child development designed to show the proportions of children in each neighbourhood who are "bouncing" or "slipping" between

kindergarten and Grade 4 (Lloyd & Hertzman, 2009). Second, for those conducting population-based development research, is that individual EDI scores are valid to be used either outcome or explanatory variables in multilevel developmental models. One example is of this is Carpiano et al. (in press), who examined the effects of family and neighbourhood affluence on school readiness.

5.11.2 Issues Relating to Relative Scoring

There are important and diverse implications for a number of stakeholders stemming from the finding of more relative teacher EDI ratings when class size is greater than 10 students. As shown above, relative scoring is a source of construct-irrelevant variance, which dilutes the appropriateness and utility of the inferences made about school readiness.

The results have implications for the designers of the EDI, in terms of considering changes to the instrument and/or its implementation in ways that would result in more absolute ratings generally. One strategy would be to reduce the number of EDI items in each scale, while maintaining adequate content coverage. Teachers would then need to focus on fewer behaviours and skills, which may improve the absolute nature of the ratings, especially if teachers get early and effective EDI training. Indeed, shorter versions of the EDI have been developed for use internationally (Janus et al., 2007), in recognition of both the larger class sizes in many countries, and the generally high response burden teachers experience for each completed EDI form.

The EDI designers may also wish to explore whether the general tendency towards relative scoring could be affected by the descriptive headings for the response categories, in particular for the three-category items. For the EDI, these category headings are either "never"/"sometimes"/"often," or "poor-very poor"/"average"/"good-very good." It is interesting to note that for the 56 items using the first descriptor type, the mean proportion of teachers

endorsing the middle category was 25%. In contrast, for the 14 items using the second descriptor type, the mean proportion endorsing the middle category was 40%. Perhaps one of the reasons for this difference is that the descriptors in the second heading type have much more of a connotation of relativity, particularly with the middle category of "average."

One potential implication for organizations that administer the EDI (e.g., HELP, the Offord Centre for Child Studies) is in the realm of EDI teacher training. Perhaps teachers would make more absolute ratings if training occurred at the beginning of the school year, rather than closer to when teachers rate the children. This would give teachers more time to observe their students on the 103 behaviours and skills addressed in the EDI items. Also, if the EDI training manual provides concrete examples of how to score each item, ratings should tend to be more absolute. The accessibility of the manual during rating is also important, so that teachers can easily consult it when unsure about how to rate a particular child. The recent adoption in British Columbia of an electronic version of the EDI helps in this regard, as specific assistance for each item is now just a matter of a mouse click on that item. The EDI training sessions and manual should also explicitly emphasize the importance of avoiding relative scoring, to make teachers more aware of this issue.

An implication for government stakeholders is that more absolute scoring, and therefore better school readiness inferences, would be expected if provincial governments had a policy to restrict kindergarten class sizes, ideally to a maximum of 10 children. This would presumably be achieved by increasing the number of kindergarten teachers and classrooms. Smaller class sizes would also improve the quality of student-teacher interactions (e.g., Blatchford et al., 2003), an even more important benefit. It would also allow teachers to more easily adopt teacher-led assessment practices with proven efficacy, such as the Work Sampling System (Meisels, 1992).

5.11.3 Issues Relating to Vulnerability Scores

It is instructive both psychometrically and practically to look for instances where actual scores and vulnerability scores showed a somewhat different pattern of EDI validation results. For example, the ICC(1) statistic for vulnerability scores, while still substantially greater than zero, was only about half the magnitude of its counterpart for actual scores; thus, classroom-level influences are smaller. In addition, the reliability of class means for vulnerability scores was moderate (range .61 to .70), again weaker than for actual scores. This is a reminder, not just for the EDI but for all multilevel measures, that psychometric properties are characteristic of scores (i.e., composition models) rather than measures. Thus, choosing appropriate composition models is a key aspect of the process and argument of validation.

The reduction in classroom-level influence and reliability for vulnerability scores reflects the fact that actual scores are rated by individual teachers, while vulnerability scores come from establishing a somewhat arbitrary cutoff score based on the collective ratings by teachers across the province. In the future, vulnerability scores could likely be made more reliable by taking advantage of standard-setting methodology developed by educational measurement specialists (see, for example, Cizek & Bunch, 2007). This could involve, for instance, consulting with kindergarten teachers and developmental psychologists about their definitions of vulnerability. Such standard setting will also be useful in the policy arena, to ensure that strategic goals, such as the B.C. provincial government's goal of reducing the proportion of developmentally vulnerable children to under 15% by the year 2015 (Province of British Columbia, 2008), are based on more absolute standards than is currently the case.

5.12 Limitations and Future Directions

One limitation of the study is that the Wave 2 data were analyzed using only the first two years of the three years comprising the wave. Therefore, the results are not based on a representative sample of kindergarten children in British Columbia – they are only representative of the 44 school districts for which data were available at the time the analyses were done. Ideally, data from a complete wave (or even multiple waves) should be used. This should not affect the pattern of results much, however, given the large size of the sample already, and the fact that the "missing" school districts are not from any one large region of the province, or are otherwise obviously systematically related to any variables at some aggregate level, such as teacher education, class size, etc.

The first of Chen et al.'s (2004a) five steps –defining the construct across levels of analysis – was not addressed empirically in this dissertation. The analysis began with the second step of articulating appropriate compositional models for aggregating individual scores. Given the previous discussion about the conceptual/philosophical complexities about the construct, omitting this first step is an important limitation. The multilevel validation procedures used in steps 2 through 5 have been applied using the current construct definitions, which may very well be seriously flawed. All results, therefore, are contingent on the assumption of current construct definitions.

While an in-depth examination of the EDI theoretical definition is needed from a psychometric standpoint, there were practical reasons for assuming the current definition in this dissertation. The most important is that the developers of the EDI, who retain control over its use, have a commitment to using the instrument as currently constructed. Much of this has to do with the large number of children so far for whom EDI scores have been collected. A new and psychometrically improved EDI would not necessarily be compatible with the current version,

rendering all current data incomparable with future data based on a newly derived instrument. There is some existing history of changes, which have created issues of comparability. On the basis of the results of a Rasch analysis of data from kindergarten children in Australia (Andrich & Styles, 2004), the number of response options in many of the original EDI items was reduced in all versions of the instrument from 2004 onwards. In an attempt to retain comparability, EDI data collected before the change were recoded using the newly revised categories (based on best guesses), which changed both average scores and the location of vulnerability cutoffs for both British Columbia and Ontario.

One important future task would be to conduct a systematic analysis of the items in each of the scales and subscales of the EDI, to establish whether each scale or subscale score is an index or a measure. An index is a compilation of items that address qualitatively distinct aspects of a construct. For example, clinical depression is diagnosed on the basis of a variety of symptoms, some or all of which may be present. This has implications for how the EDI should be validated, as items making up an index have lower inter-item consistency expectations than items making up a measure (Bollen & Lennox, 1991). This may help increase the apparent validity of the EDI, especially for the one PHWB subscale that has a very low coefficient alpha reliability. Doing such an analysis would also help in future revisions of the EDI. For example, it might be decided that for some of the index-like qualities of a subscale or scale, some items might be redundant (as maybe was shown in the Rasch analysis) or other items may be needed to provide good content coverage of the latent variable that the index is representing.

To date, there have been no descriptive or inferential studies that address issues relating to "don't know" and missing responses for the EDI. There have been attempts to reduce nonresponse by moving to an electronic version of the instrument, which alerts teachers at the end of each section about items they have left blank. Data are not yet available to establish

whether this change in method has had this effect. A future study could address which items are most prone to nonresponse, and perhaps test some teacher training interventions targeted to those items. Another study could try to model scale missingness, either as a categorical variable (no items missing vs. some missing vs. all missing) against a variety of available child, teacher, and classroom characteristics.

Recently, the focus of aggregate EDI construct conceptualization at HELP has been on new ways to characterize EDI scores meaningfully, particularly at the neighbourhood level of aggregation. At issue is a desire to move beyond central tendency or percent vulnerable to capture something more essential or qualitative about the diversity of neighbourhoods. In the Atlas of Child Development (Kershaw et al., 2005), five idealized neighbourhood types were proposed: high challenge, buffered, average, wide-ranging, and low challenge. These types were conceptualized in terms of various combinations of average scores and vulnerability rates, though the five types have never been operationally defined. Currently, a small validation group at HELP is moving towards operationalizing neighbourhood type using a larger and more complete set of criteria that attempt to capture what development means at the neighbourhood level. These new criteria could certainly include measures of variability of individual scores. Thus, the dispersion compositional model, with its own particular validation issues, may be relevant someday in the validation of the EDI.

A critical future direction will be the ongoing exploration of functional relationships, at various levels of aggregation, between the EDI and other variables in its nomological network. This is the final step in Chen et al.'s (2004a) multilevel validation framework, and goes to the heart of the unitary view of validity (Zumbo, 2007a, 2009). This view, which is consistent with the social constructivist underpinnings of the EDI, stresses the necessity of explaining, not just measuring, covariation between the construct of interest and other theoretically relevant

variables. Teacher and classroom influence is only one isolated (but important) aspect of our understanding of this multilevel construct.

A case in point is the exploration of the role of various neighbourhood-level socio-economic factors on EDI scores, which is well underway. These consist both of multilevel studies (e.g., Carpiano et al., in press) and single level studies using neighbourhood-level EDI scores, based both on actual scores (e.g., LaPointe, 2006) and vulnerability scores (Kershaw & Forer, 2006). Guhn and Goelman (2009) emphasize the need to collect more proximal process data (e.g., teacher observations, parenting surveys) to address an important component that is currently missing in EDI research.

From this same functional exploration viewpoint, it is also important to consider the EDI not only as an outcome measure, but also in the context of its theoretical link to constructs like school achievement that are measured later in time. The Community Index of Child Development (CICD; Lloyd & Hertzman, 2009), which measures population-level developmental trajectories from kindergarten to Grade 4, is a good example of this. Given HELP's mandate to understand the social and health determinants of developmental trajectories over the whole life course, the eventual goal is to link EDI data to outcomes throughout formal education and beyond.

Future WABA analyses would be ideal as a starting place for examining these multilevel patterns of covariation between EDI scores and other theoretically related variables. WABA is uniquely equipped to test appropriate levels of aggregation, rather than set them in advance, as is the case for HLM or SEM analyses. In addition, unlike these other techniques, WABA analyses can consider any number of levels at once. Thus, WABA and multilevel modeling techniques naturally play complementary roles in multilevel validation work.

Unfortunately for the multilevel validation of the EDI, establishing and testing the appropriate levels for relationships with theoretically relevant variables is mostly limited by the current availability of these relevant variables at the individual level. There are only a handful of child-level variables measured by teachers as part of completing the EDI (e.g., age, gender, language spoken at home, Aboriginal identity, English as a second language status); some of these are of dubious accuracy or have a lot of missing data (e.g., early childhood education and care programs attended before kindergarten). Available socioeconomic variables mostly come from Statistics Canada, and are only available at the neighbourhood and school district levels. As part of the mandate of the Human Early Learning Partnership, efforts are being made to gain secure and private access to a wide variety of individual-level data, from Ministries of Health, Education, and Children and Family Development. Once linked to EDI data, either probabilistically or through government identification numbers (such as the Personal Health Number or Personal Education Number), there will be many more theoretically relevant variables available that can be expressed at a variety of levels.

An interesting future study would be to ask teachers to rate each student as vulnerable or not, either in some overall way, or for the five developmental domains separately. The question would have to be carefully phrased, asking them to consider their current class against all kindergarten children they have ever taught, to avoid class-specific relative vulnerability classifications. This would help in two ways: first, to compare provincial or national cutoffs that have been arbitrarily established with teachers' absolute assessments of individual vulnerability, and second, to establish the extent to which teachers vary in their own definitions of "vulnerable," based on the associated scale scores.

References

Allison P. D. (2002). *Missing data*. London : Sage.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2, Pt. 2).

Andrich, D., & Styles, I. (2004). *Report on the psychometric properties of the Early Development Instrument (EDI) using the Rasch model.* A technical paper commissioned for the development of the Australian Early Development Instrument (AEDI), Murdoch University.

Beswick, J. F., Sloat, E. A., & Willms, J. D. (2004). A comparative study of bias in teacher ratings of language and emergent literacy skills. Unpublished manuscript.

Blatchford, P., Bassett, P., Goldstein, H., & Martin, C. (2003). Are class size differences related to pupils' educational progress and classroom processes? Findings from the Institute of Education Class Size Study of Children Aged 5 to 7 Years Old, *British Educational Research Journal, 29*, 709-730.

Blatchford, P., Russell, A., Bassett, P., Brown, P., & Martin, C. (2007). The effect of class size on the teaching of pupils aged 7 – 11 years old. *School Effectiveness and School Improvement, 18*, 147-172.

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

Bliese, P. D., & Halverson, R. R. (1998). Group size and measures of group-level properties: An examination of eta-squared and ICC values. *Journal of Management, 24*, 157-172.

Bollen, K. & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305-314.

Brinkman, S., Silburn, S., Lawrence, D., Goldfeld, S., Sayers, M., & Oberklaid, F. (2007). Investigating the validity of the Australian Early Development Index. *Early Education and Development, 18*, 427-51.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32*, 513-531.

Bronfenbrenner, U., & Morris, P. A. (1998).The ecology of developmental processes. In *W. Damon (Series Ed.) & R. M. Lerner (Vol. Ed.), Handbook of child psychology: Vol. 1: Theoretical models of human development* (pp. 993-1028).New York: Wiley.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.

Carlton, M. P., & Winsler, A. (1999). School readiness: The need for a paradigm shift. *School Psychology Review, 28,* 338-352.

Carpiano, R. M., Lloyd, J. E. V., & Hertzman, C. (in press). Concentrated affluence, concentrated disadvantage, and children's readiness for school: A population-based, multi-level investigation. *Social Science & Medicine*.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83,* 234-246.

Chen, G., Mathieu, J. E., & Bliese, P. D. (2004a). A framework for conducting multi-level construct validation. *In F. J. Yammarino & F. Dansereau (Eds.). Multi-level issues in organizational behaviour and processes*. (pp. 273-303) Elsevier: The Netherlands.

Chen, G., Mathieu, J. E., & Bliese, P. D. (2004b). Validating frogs and ponds in multi-level contexts: Some afterthoughts. *In F. J. Yammarino & F. Dansereau (Eds.). Multi-level issues in organizational behaviour and processes*. (pp. 335-343) Elsevier: The Netherlands.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage Publications, Inc..

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.).Mahwah, NJ: Erlbaum.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Dansereau, F., Alutto, J. A., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The "varient" approach*. Englewood Cliffs, NJ: Prentice-Hall.

Dansereau, F., Cho, J., & Yammarino, F. J. (2006). Avoiding the fallacy of the wrong level. *Group and Organization Management, 31,* 536-577.

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

Dansereau, F., & Yammarino, F. J. (2004). Overview: Multi-level issues in organizational behavior and processes. *In F. J. Yammarino & F. Dansereau (Eds.). Multi-level issues in organizational behaviour and processes*, (pp. xiii-xxxiii) Elsevier: The Netherlands.

Dansereau, F., & Yammarino, F. J. (2006). Is more discussion about levels of analysis really necessary? When is such discussion sufficient? *The Leadership Quarterly, 17*, 537-552.

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

Doherty, G. (1997). *Zero to six: The basis for school readiness* (No. R-97-8E). Ottawa, ON: Human Resources Development Canada, Applied Research Branch, Strategic Policy.

Duku, E., & Janus, M. (April, 2004). Stability and reliability of the Early Development Instrument: A population-based measure for communities (EDI). Department of Psychiatry and Biobehavioural Sciences, McMaster University, Annual Research Day.

Duncan, G.J., A. Claessens, A.C. Huston, L.S. Pagani, M. Engel, H. Sexton, C.J. Dowsett et al. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428-46.

Ellwein. M. C., Walsh, D. J., Eads, G. M., & Miller, A. (1991). Using readiness tests to route kindergarten students: The snarled intersection of psychometrics, policy, and practice. *Educational Evaluation and Policy Analysis, 13*, 159-175.

Englehard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261-288), Mahwah, NJ: Erlbaum.

Epstein, J. N., Willoughby, M., Valencia, E. Y., Abikoff, H. B., Arnold, L. E., & Hinshaw, S. P. (2005). The role of children's ethnicity in the relationship between teacher ratings of attention-deficit/hyperactivity disorder and observed classroom behavior. *Journal of Consulting and Clinical Psychology, 73*, 424-434.

Finn, J. D., Pannozzo, G. M., & Achilles, C M. (2003). The "whys" of class size : Student behavior in small classes. *Review of Educational Research, 73*, 321-368.

Ford, D. L., & Lerner, R. M. (1992). *Developmental systems theory: An integrative approach*. Newbury Park, CA: Sage.

Forget-Dubois, N., J.-P. Lemelin, M. Boivin, D. Ginette, J.R. Séguin, F. Vitaro, et al. (2007). Predicting early school achievement with the EDI: A longitudinal population-based study. *Early Education and Development, 18*, 405-26.

Gesell, A. (1925). *The mental growth of the preschool child*. New York: Macmillan.

Gorsuch, R. L. (1983). Factor Analysis, 2nd ed. Hillsdale, NJ: Erhlbaum.

Gnezda, M. T., & Bolig, R. (1988, November). *A national survey of public school testing of pre-kindergarten and kindergarten children*. Washington, DC: National Forum on the Future of Children and Families, National Research Council.

Graue, E. (2006). The answer is readiness – Now what is the question? *Early Education and Development, 17*, 43-56

Griffith, J. (2002). Is quality/effectiveness an empirically demonstrable school attribute? Statistical aids for determining appropriate levels of analysis. *School Effectiveness and School Improvement, 13*, 91-122.

Grilli, L., & Rampichini, C. (2007). Multilevel factor models for ordinal variables. *Structural Equation Modeling, 14*, 1-25.

Guhn, M., Gadermann, A., & Zumbo, B. D. (2007). Does the EDI measure school readiness in the same way across different groups of children? *Early Education and Development, 18,* 453-472.

Guhn, M. & Goelman, H. (2009). A bioecological approach to theory and validity of a population-based child development measure. *Social Indicators Research*. Manuscript under review.

Guhn, M., Janus, M., & Hertzman, C. (2007). The Early Development Instrument: Translating school readiness assessment into community actions and policy planning. *Early Education and Development, 18,* 369-374.

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). Elsevier: The Netherlands.

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Erlbaum.

Hu, L. T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.

James, L.R., Demaree, R.G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85-98.

Janus, M. (May, 2001). *Validation of a teacher measure of school readiness with parent and child-care provider reports*. Department of Psychiatry and Biobehavioural Sciences, McMaster University, Annual Research Day.

Janus, M. (April, 2002). *Validation of the Early Development Instrument in a sample of First Nations children*. Department of Psychiatry and Biobehavioural Sciences, McMaster University, Annual Research Day.

Janus, M., Brinkman, S., Duku, E., Hertzman, C., Santos, R., Sayers, M. et al. (2007). *The Early Development Instrument: A population-based measure for communities. A handbook on development, properties, and use*. Hamilton, ON: Offord Centre for Child Studies.

Janus, M., & Duku, E. (2007). The school entry gap: Socioeconomic, family, and health factors associated with children's school readiness to learn. *Early Education and Development, 18*, 375-403.

Janus, M., Harren, T., & Duku, E. (2004, April). *Neighbourhood perspective on school readiness in kindergarten, academic testing in Grade 3, and affluence levels*. Paper presented at the McMaster University Psychiatry Research Day, Hamilton, ON.

Janus, M., & Offord, D. (2000). Readiness to learn at school. *ISUMA, 1*, 71-75.

Janus, M., & Offord, D. (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.

Janus, M., Offord, D., & Walsh, C. (April, 2001). *Population-level assessment of readiness to learn at school for 5-year-olds in Canada: Relation to child and parent measures*. Presented at the Society for Research on Child Development conference, Minneapolis.

Janus, M., Walsh, C., & Duku, E. (2005, March). *Early Development Instrument: Factor structure, sub-domains and Multiple Challenge Index.* Department of Psychiatry and Biobehavioural Sciences, McMaster University, Annual Research Day

Janus, M., Walsh, C., Viveiros, H., Duku, E., & Offord, D. (2003, April). *School readiness to learn and neighbourhood characteristics*. Poster presented at the Biennial meeting of the Society for Research on Child Development, Tampa, FL.

Janus, M., Willms, J. D., & Offord, D. R. (2000). *Psychometric properties of the Early Development Instrument (EDI): A teacher-completed measure of children's readiness to learn at school entry*. Unpublished manuscript.

Keating, D. P. (2007). Formative evaluation of the Early Development Instrument: Progress and prospects. *Early Education and Development, 18(3),* 561-570.

Keating, D. P., & Hertzman, C. (Eds.). (1999). *Developmental health and the wealth of nations*. New York: Guildford.

Kershaw, P., & Forer, B. (2006, April). *What are the social and economic indicators of nurturing neighborhoods?* Poster presented at the American Educational Research Association conference, San Francisco, California.

Kershaw, P., Irwin, L., Trafford, K., & Hertzman, C. (2005). *The British Columbia Atlas of Child Development*. Human Early Learning Partnership: Vancouver, BC.

Kim, J., & Suen, H. K.(2003). Predicting children's academic achievement from early assessment scores: A validity generalization study. *Early Childhood Research Quarterly,* 18, 547-566.

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). Elsevier: The Netherlands.

Klein, K. J., Bliese, P. D., Kozlowski, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A., et al. (2000). Multilevel analytic techniques: Commonalities, differences, and continuing questions. In K. J. Klein & S. W. J. Kozlowski (Eds.). *Multilevel theory, research, and methods in organizations* (pp. 512-553). San Francisco: Jossey-Bass.

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.

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* (pp. 3-90). San Francisco: Jossey-Bass.

Kreiger, J. (February, 2003). *Class size reduction: Implementation and solutions.* Paper presented at the SERVE Research and Policy Class Size Symposium, Raleigh, SC.

LaPointe, V. R. (2006). *Conceptualizing and examining the impact of neighbourhoods on the school readiness of kindergarten children in British Columbia*. Unpublished Doctoral Dissertation, Vancouver: The University of British Columbia.

LaPointe, V. R., Ford, L., & Zumbo, B. D. (2007). Examining the relationship between neighbourhood environment and school readiness for kindergarten children. *Early Education and Development, 18*, 473-496.

Lloyd, J. E. V., & Hertzman, C. (2009) From Kindergarten readiness to fourth-grade assessment: Longitudinal analysis with linked population data. *Social Science & Medicine, 68*, 111-123.

Love, J. M., Aber, J. L., & Brooks-Gunn, J. (1994). *Strategies for assessing community progress towards achieving the First National Educational Goal*. Princeton, NJ: Mathematica Policy Research.

Mashburn, A. J., & Henry, G. T. (2004). Assessing school readiness: Validity and bias in preschool and teachers' ratings. *Educational Measurement: Issues and Practices*, *23*, 16-30.

Mashburn, A. J., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2006). Teacher and classroom characteristics associated with teachers' ratings of prekindergartners' relationships and behaviors. *Journal of Psychoeducational Assessment, 24*, 367-380.

Meisels, S. J. (1992). *The Work Sampling System*. Ann Arbor, MI: Center for Human Growth and Development, University of Michigan.

Meisels, S. J. (1999). Assessing readiness. In R. C. Pianta & M. J. Cox (Eds.), *The transition to kindergarten* (pp. 39-66), Baltimore, MD: Paul H. Brookes.

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

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Muthen, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22,* 276-398.

Muthen, L. K., & Muthen, B. O. (2006). *Mplus User's Guide, version 4*. Los Angeles: Muthen and Muthen.

National Education Goals Panel. (1991). *The National Education Goals report, 1991: Building a nation of learners*. Washington, DC: U.S. Government Printing Office.

National Research Council Institute of Medicine. (2000). From neurons to neighbourhoods: The science of early childhood development. In *Committee on Integrating the Science of Early Childhood Development*, J. P. Shonkoff & D. A. Phillips (Eds.) (pp. 328-336). Washington, D.C.: National Academy Press.

Neal, L. I., McCray, A. D., Webb-Johnson, G., & Bridgest, S. T. (2003). The effects of African-American movement styles on teachers' perceptions and reactions. *Journal of Special Education, 37,* 49-57.

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

Ostroff, C. (1992). The relationship between satisfaction, attitudes and performance: An organizational-level analysis. *Journal of Applied Psychology, 77,* 963-974.

Pedder, D. (2006). Are small classes better? Understanding relationships between class size, classroom processes, and pupils' learning. *Oxford Review of Education, 32,* 213-234.

Pellegrini, A. and Blatchford, P. (2000) *The child at school: Interactions with peers and teachers*. London: Edward Arnold.

Province of British Columbia (2008). *Strategic plan 2008/09 – 2010/11*. Retrieved from http://www.bcbudget.gov.bc.ca/2008/stplan/2008_Strategic_Plan.pdf on September 18, 2008

Rimm-Kaufman, S. E., & Pianta, R. C. (2000). An ecological perspective on the transition to kindergarten: A theoretical framework to guide empirical research. *Journal of Applied Developmental Psychology, 21,* 491-511.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351-357.

Rosenthal, R., & Jacobson, L. (1992). *Pygmalion in the classroom (expanded ed.)*. New York: Irvington.

Rowan, B., Raudenbush, S. W., & Sang, J. K. (1991). School climate in secondary schools. In Raudenbush, S. W. & Willms, J. D. (Eds.), *Schools, Classrooms, and Pupils: International studies of schooling from a multilevel perspective* (pp. xx). San Diego: Academic.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality rating data. *Psychological Bulletin, 88*, 413-428.

Saluja, G., Scott-Little, C., & Clifford, R. M. (2000). Readiness for school: A survey of state policies and definitions. *Early Childhood Research & Practice, 2*. Retrieved February 12, 2007 from http://ecrp.uiuc.edu/v2n2/saluja.html

Satorra, A., & Bentler, P. M. (1999). *A scaled difference Chi-square test statistic for moment structure analysis.* Technical Report, University of California, Los Angeles. http://preprints.stat.ucla.edu/260/chisquare.pdf

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Serwatka, T. S., Dove, T., & Hodge, W. (1986). Black students in special education: Issues and implications for community involvement. *Negro Education Review, 37*, 17-26.

Shepard, L. A. (1997). Children not ready to learn? The invalidity of school readiness testing. *Psychology in the Schools, 34*, 85-97.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Snow, K. L. (2006). Measuring school readiness: Conceptual and practical considerations. *Early Education and Development, 17*, 7-41.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th edition). New York: Routledge.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds. & Trans.). Cambridge, MA: Harvard University Press.

Zumbo, B.D. (2007a). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics*, (pp. 45-79). Elsevier Science B.V.: The Netherlands.

Zumbo, B. D. (2007b). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4,* 223-233.

Zumbo, B. D. (2009). Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice. In Robert W. Lissitz (Ed.) *The Concept of Validity: Revisions, New Directions and Application,* (pp. 65-82). IAP - Information Age Publishing, Inc.: Charlotte, NC.

Zumbo, B. D. & Forer, B. (in press). Testing and measurement from a multilevel view: Psychometrics and validation. In Bovaird, J., Geisinger, K., & Buckendahl, C. (Eds.). *High Stakes Testing in Education – Science and Practice in K-12 Settings (Festschrift to Barbara Plake).* American Psychological Press, Washington, D.C.

Appendix A

Steps to Calculate of the Root Mean Square Residual – Proportion (RMR-P) Fit Statistic

The calculation of the RMR-P follows the same basic procedural steps used to calculate the Standardized Root Mean Square Residual fit statistic (SRMR; Hu & Bentler, 1995) which is based on residual covariances.

1. The calculation begins with the residual proportions that result from conducting a CFA with categorical items. For each pair of items, there will be c_1*c_2 residual proportions, where c_1 and c_2 are the number of categories in items 1 and 2. For example, for two items each with three categories, there will be nine residual proportions per item pair. For a domain with x items, there are $x(x-1)/2$ item pairs. Therefore, for the emotional maturity scale for example, which has 30 items each with three categories, there will be 435 item pairs each with nine residual proportions, for a total of 3,915 data points.
2. Since the residual proportions for any item pair must necessarily add up to zero, the relative size of the residuals was assessed using the strategy of squaring and later taking the square root. Therefore, the next step was to square all of the residual proportions.
3. The mean of all the squared residuals was then calculated.
4. Finally, the square root of the mean squared residuals was taken, resulting in the unscaled RMR-P statistic.

5. The theoretical maximum value of the unscaled RMR-P depends on the number of categories in each item (as shown below). Because some items are binary and others are three-category

ordinal, a scaling factor needs to be applied to set a common scale for the RMR-P for each domain. The calculation of this scaling factor is as follows:

a. For any item pair, regardless of the number of categories, the maximum sum of the squared residual proportions is 2. This follows mathematically from the observations that the sum of all observed proportions must equal 1, and the sum of all residual proportions must equal 0. The poorest possible model fit is when one residual proportion is equal to 1 and a second residual proportion is equal to (-1), and the rest of the residuals are equal to 0. In this case, the sum of the squared residual proportions is equal to 2 [i.e., $1^2 + (-1)^2$]. Any other combination of possible residual proportions will result in a lower sum of squares.
b. For a domain with only binary items (i.e., language and cognitive development), the maximum RMR-P statistic is 0.71. This is because in step 3, the maximum mean of the squared residuals would be 0.5 (maximum of 2 for each four squared residual proportions); the square root of 0.5 (step 4) is 0.71. For a domain with only three-category items (i.e., social competence, emotional maturity, and communication and general knowledge), the maximum unscaled RMR-P is 0.47, based on a maximum mean of the squared residuals of 0.22 (maximum of 2 for each nine squared residual proportions); the square root of 0.22 is 0.47.
c. The only EDI domain with a mixture of binary and three-category items is physical health and well-being. The maximum RMR-P for this domain is .59, based on weighting the number of item pairs with four (2x2), six (2x3), and nine (3x3) residual proportions.
d. Having calculated the maximum RMR-P for each domain, these maxima can then be used as scaling factors, so that for all domains, the range of the RMR-P statistic is from 0 to 1. Therefore, the unscaled RMR-P statistics resulting from step 4 are divided by the appropriate scaling factor for each domain.

e. These final RMR-P statistics are now on the same scale used for the SRMR. This certainly does not mean that RMR-P scores are in any way standardized, but it does result in a metric with the same range as the SRMR. On the basis of this similar metric, and given that the RMSR fit statistic is itself not based on any theoretical distribution, the same conventional .05 cutoff is then applied to the scaled RMR-P statistic when assessing model fit.