Alberta Journal of Educational Research, Vol. 62.3, Fall 2016, 288-305

Measurement Invariance of Early Development Instrument (EDI) Domain Scores Across Gender and ESL Status

Amin Mousavi[1], Vijaya Krishnan[2]

[1]University of Saskatchewan, [2]University of Alberta

The Early Development Instrument (EDI) is a widely used teacher rating tool to assess kindergartners' developmental outcomes in Canada and a number of other countries. This paper examines the measurement invariance of EDI domains across ESL status and gender by means of multi-group confirmatory factor analysis. The results suggest evidence of measurement invariance for physical health and well-being, social competence, emotional maturity and language and cognitive development domains. Moreover, the communication skills and general knowledge domain did not show acceptable fit in terms of RMSEA. The results and potential explanations are discussed.

L'instrument de mesure du développement de la petite enfance (IMDPE) est un outil d'évaluation largement utilisée pour mesurer le développement des élèves en maternelle au Canada et dans d'autres pays. Cet article porte sur l'équivalence de mesure des domaines de l'IMDPE entre le statut d'ALS et le sexe par une analyse factorielle confirmatoire multigroupe. Les résultats font ressortir des preuves d'équivalence de mesure pour les domaines de la santé physique et le bienêtre, la compétence sociale, la maturité affective, le développement langagier et cognitif. De plus, le domaine des compétences en communication et des connaissances générales n'a pas démontré une correspondance acceptable par rapport à l'erreur quadratique moyenne de l'approximation (RMSEA). Nous discutons des résultats et proposons des explications possibles.

The Early Development Instrument (EDI) is a tool to assess kindergarteners' development in the five areas: physical health and well-being, social competence, emotional maturity, language and thinking skills, and communication and general knowledge. The tool is designed to be universal enough to be relevant to most preschoolers around the world, regardless of their racial or ethnic background. The EDI results will enable communities to strengthen the qualities of our programs by focussing on something that we all know is very important to children's overall health and well-being—developmental health. The tool is geared to provide a methodology and a framework for communities to effectively address developmental difficulties in children at a macro-level. Specifically, the EDI is a survey-based thematic tool primarily designed to assist and target communities at a local level, although data are collected through surveys. The EDI provides an assessment of the five areas with no component of screening, yet constructed from the perspective of a Eurocentric epistemology (Krishnan, 2013). A legitimate concern, then, is that it may in fact be neglecting crucial areas of development among children of different ethnic

backgrounds.

Data drawn from such surveys as the EDI can facilitate and encourage community monitoring of the developmental health of our young children, predict learning and behaviour, and mobilize stakeholders towards positive change in local areas. The test scores can serve many purposes, including the categorization of children into groups by linking them to cut scores and investigation of relationships between variables within groups or group comparisons, mostly with the application of ANOVA. The underlying assumption is that scores are comparable across groups, with no violation of measurement invariance. From a measurement perspective, scores can only be comparable if an instrument measures an underlying construct in the same manner across groups. This property is referred to as "Measurement Invariance (MI)" (Millsap, 2011). As Schmitt and Kuljanin (2008) put it,

> [A] measure is invariant when members of different populations who have the same standing on the construct being measured receive the same observed score on the test. A test violates invariance when two individuals from different populations who are identical on the construct score differently on it. (p. 211)

The EDI has been validated by over ten years of research in different educational settings at either the junior or senior kindergarten level with no clear disparities in the performances of children across different groups. In other words, it is supposed to be invariant across groups providing comparable scores for children, roughly among four and seven year olds. Indeed, the assumption is that the tool can very well be used in order to draw similar inferences within any socio-cultural settings. However, the utility of the instrument is, of course, very much dependent upon its strength to generalize across population groups, whether it be age, sex, or any other social and demographic background. Given its widespread use and potential applications, it is important to thoroughly understand the characteristics of EDI test scores.

Despite its importance, to our knowledge, there are only two studies that examined MI of EDI. Guhn, Gadermann, and Zumbo (2007) analyzed MI of EDI items that utilized differential item functioning (DIF). In DIF, the aim is to evaluate whether items function the same across groups, and it can also be extended to test scores referred to as differential test functioning (DTF). Guhn et al. (2007) employed ordinal logistic regression in order to assess differential functioning of EDI, across gender, English as a Second Language (ESL) status, and Aboriginal status. They found no systematic MI on gender and Aboriginal status, except for one item on which boys were found more likely to be rated as physically aggressive by Kindergarten teachers. By contrast, ESL children systematically received lower ratings on items of the language and communication domain. The authors also reported that DIF at domain score level was quite substantial. Their conclusion was that since EDI is a community-level measure and observed DIF on some items was expected, the observed DIF is attributable to item impact rather than item bias. A point worth mentioning is that the authors utilized the DIF method in their study to explore group differences at the domain score level (i.e. total observed score), which implies that score equivalence was assumed. However, as some would suggest (e.g., Schmitt & Kuljanian, 2008; Steinmetz, 2013; Vandenberg & Lance, 2000), score equivalency requires scalar invariance for group comparisons using observed total (or mean) scores. So, any conclusion on group differences based on observed scores without providing evidence of scalar invariance can be sceptical. In another study, Duku, Janus and Brinkman (2014) applied multi-group confirmatory factor analysis (MGCFA) for categorical data in order to evaluate MI of the short

form of EDI (Janus & Duku, 2005) across gender and two Asian countries (Philippines and Indonesia). Based on their results, the social competence and emotional maturity domains showed an acceptable level of MI across gender and countries. One of the main limitations of their study is that the short form of EDI has not been published publicly, and there is no information about the items and configuration of the short form of EDI.

One of the most important principles to remember is that the EDI scores have limited meaning unless the outcome measures are the same or very close to one another, across population groups. To measure children's development, independent of demographic variations, items on the EDI need to be free from biases. More research on the outcome gaps, if any, will be useful because of the important implications for interventions. No matter how small the extent of vulnerability or developmental delays, it is crucial to address the needs and priorities of different groups, clarifying the relevance of variations in developmental patterns in conjunction with age, sex, ethnicity, and so on. This research aims to provide insights into one aspect of MI of the original form of EDI. That is, the question addressed is: are the EDI outcome-measures performed in the same manner across ESL status and the gender of children?

The Concept of Measurement Invariance

In social and behavioural sciences, most of the variables of interest, if not all, are unobservable. These unobserved or latent variables can be measured indirectly via measurement models (Borsboom, Mellenbergh, & van Heerden, 2003). This procedure entails levels of uncertainty and error of measurement. So, it is important to check whether items are interpreted in the same way by respondents; scales are calibrated consistently, and observed mean differences mirror latent mean differences across groups (Drasgow, 1984). More generally, the study of measurement equivalence of a tool across groups is called measurement invariance. In latent variable modelling, the relationship between an observed score X_i and an underlying construct can be defined as:

$$X_i = \lambda_i \xi + \delta_i \quad (1)$$

Where, ξ is latent variable, λ_i is corresponding factor loading and δ_i represents error of measurement. Accordingly, an observed response such as X_i is influenced by latent variable, factor loading, and error of measurement. Equation 1 is the most common form of relationship between an observed score and a latent variable in analyzing the variance-covariance matrix of observed data. The variance-covariance matrix of observed variables can be obtained by taking the variance-covariance of the same equation as:

$$\Sigma_{xx} = \Lambda_x \Phi \Lambda'_x + \Psi \quad (2)$$

In equation 2, Σ_{xx} is the variance-covariance matrix of observed data, Λ_x is the vector of factor loading for all observed variables, Φ denotes the correlation matrix between latent variables, and Ψ represents the variance-covariance matrix between measurement errors. As evident in equations 1 and 2, the mean of observed responses has not been considered because the focus of such an analysis is on the variance-covariance matrix rather than on mean differences. When considering observed and latent means in analysis, equation 1 can be extended as:

$$X_i = \tau_i + \lambda_i \xi + \delta_i \quad (3)$$

Where, τ_i is the intercept of the observed variable, X_i. The same equation can be expressed in terms of expected values, as in equation 4:

$$E(X_i) = \tau_i + \lambda_i E(\xi) + E(\delta_i) \quad (4)$$

If we assume the error of measurement to be distributed with a mean of zero, we can eliminate the last term and instead use κ for latent mean. Equation 4 can thus be reduced to:

$$E(X_i) = \tau_i + \lambda_i \kappa \quad (5)$$

Equation 5 states that the mean of the observed variable, $E(X_i)$ is a function of latent mean which is being weighted by factor loading and the intercept of the observed variable. Equations 2 and 5 can be generalized to multiple samples as:

$$\Sigma_{xx}^{(g)} = \Lambda_x^{(g)} \Phi^{(g)} \Lambda'^{(g)}_x + \Psi^{(g)} \quad (6)$$

$$E(X_i)^{(g)} = \tau_i^{(g)} + \lambda_i^{(g)} \kappa^{(g)} \quad (7)$$

Where, g represents the g^{th} group of G populations. The importance of MI as a pre-requisite for any measurement instrument has been widely recognized (e.g., Schmitt & Kuljanian, 2008; Vandenberg & Lance, 2000). More specifically, equation 6 implies the equivalency of factorial structure across groups which can be used for correlation-based analysis whereas equation 7 implies the equivalency of mean structure across groups which can be used for studies of mean differences.

If an instrument is assumed to measure a latent variable like ξ, non-invariantly across groups, then equations 6 and 7 should not be significantly different across g groups. Generally speaking, there are three important types of MI (van de Schoot, Lugtig, & Hox, 2012):

1. *Configural invariance* which tests whether or not the same pattern of fixed and free parameters holds across groups. Figure 1 illustrates a simple example where a factor with three indicators in which X_1 to X_3 are items, λ_1 to λ_3 are factor loadings and τ_1 to τ_3 are item intercepts. The error terms are not shown for the sake of simplicity. Configural invariance implies that everything is the same across both the groups in this example.
2. *Metric invariance* where factor loadings are held equal across groups but intercepts are allowed to be different. This model tests whether the construct of interest has the same meaning across groups or not. This is also called *weak factorial invariance* (Byrne & van de Vijver, 2010). Using Figure 1, a metric invariance implies that only λ_1 to λ_3 are held equal between the two groups and all other parameters are estimated for each group.

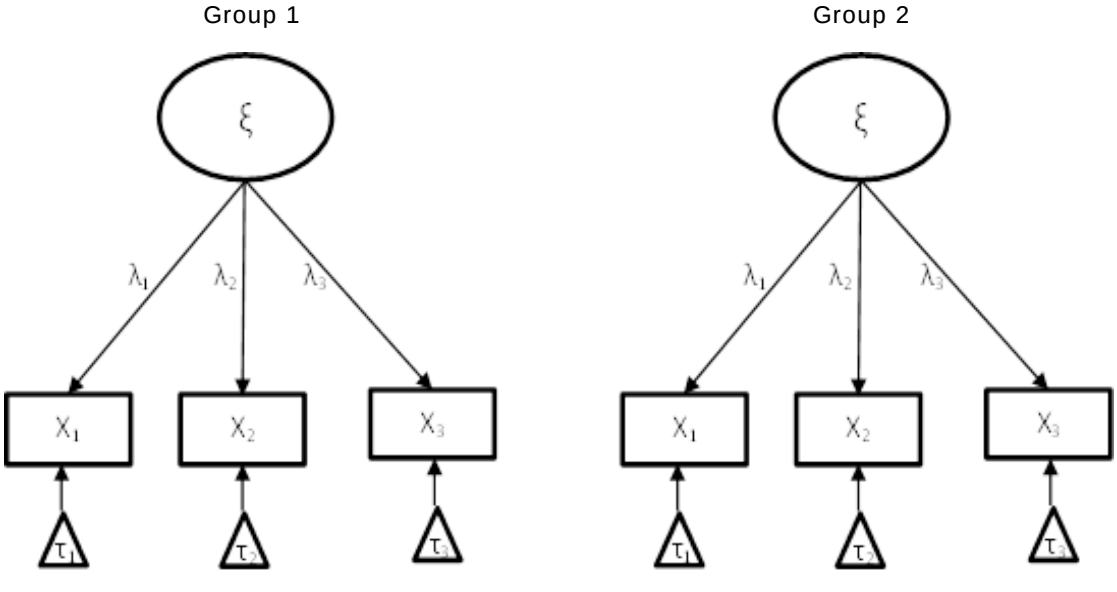

Figure 1. Illustration of configural invariance across two groups

3. *Scalar invariance* in which both factor loadings and intercepts are held equal across groups. This model allows evaluating latent mean differences across groups. This is also called *strong factorial invariance* (Byrne & van de Vijver, 2010). Using Figure 1, scalar invariance implies that λ_1 to λ_3 in addition to τ_1 to τ_3 are held equal between the two groups and all other parameters are estimated for each group.

There are two approaches in testing MI. Authors like Millsap (2011) suggested starting from a configural model and proceeding to more restricted models whereas others like Muthén and Asparhouhov (2002) suggested the reverse. Both approaches should result in the same conclusion but the second approach seems more practical because if one can determine a strong factorial invariance across groups, then the test for other types of MI is not necessary.

Although MI may have other advantages, our reasons for its application can be summarized under two ideas: (1) the psychometric evaluation of EDI. That is, to find answers to such questions as: is EDI as an outcome variable equivalent across boys and girls? And (2) cross-cultural comparison of performance as measured by EDI outcomes of kindergartners using such proxies as ESL status.

Method

Data

Our study of a cohort of children aged 4 to 7 years in the EDI is a part of a large database collected across years, 2009-2013. For Alberta, the first EDI collection was undertaken in 2009, and more school authorities were able to join the collection process in later years. The last wave of data collection occurred in 2013. As indicated earlier, the data are collected for individual children, but the results are always reported at a group level (provincial, community, sub-community, or school district). Kindergarten teachers complete the EDI questionnaire, which is made up of 103 items or questions, for each child in their class based on their observations and knowledge of each child. The child is not present, and although parents do not have to complete the questionnaire, in Alberta, they need to sign a consent form and return it to the school for the child to be included in the survey. More specifically, the EDI was not completed for children whose parent/caregiver did not sign a consent form, and in situations where the teacher had known a child for less than a month, his/her questionnaires was not used for analysis even if it was completed for most parts.

Sample

The sample consisted of 15,921 children (male: 50.8%; female: 49.2%), collected in 2013 in Alberta, Canada. This sample did not include children with special needs. The children ranged in age from 4 years to 6 years and 11 months, with a mean age of 5 years and 8 months. The sample was divided into four age groups: 5 years and 2 months and below, 5 years and 3 months to 5 years and 6 months, 5 years and 7 months to 5 years and 10 months, and 5 years and 11 months and above, and boys and girls.

Measure

Domains and sub-domains. As noted earlier, EDI is a teacher-completed instrument

comprised of 103 items, developed in Canada by Janus and Offord (2007) and includes five major domains of child development: physical health and well-being (PHYS), social competence (SOC), emotional maturity (EMOT), language and cognitive development (LANG), and communication skills and general knowledge (COMM). The items are distributed within five domains: 13 items for physical health and well-being, 26 items for social competence, 30 items for emotional maturity, 26 items for language and cognitive development, and 8 items for communication skills and general knowledge. Thirty three items are dichotomously scored items (e.g., yes/no) and 70 items are polytomously scored items (e.g., poor/very poor; average; very good/good). The first four domains also contain sub-domains: the physical health and well-being domain consists of three sub-domains (preparedness for school day, physical independence, and gross and fine motor skills); the social competence domain includes four sub-domains (overall social competence, responsibility and respect, approaches to learning, and readiness to explore new things); the emotional maturity domain consists of four sub-domains (pro-social and helpful behaviour, anxious and fearful behaviour, aggressive behaviour, and hyperactive behavior) and the language and cognitive development domain includes four sub-domains (basic literacy, advanced literacy, interest and memory, and basic numeracy) (Janus, Brinkman & Duku, 2011).

Vulnerability, the derived measure. As noted earlier, each item on the EDI is either dichotomous (yes/no) or polytomous (e.g., very good/good, average and poor/very poor). All in all, the Alberta version of the instrument currently in use has a mix of dichotomous and polytomous items.[1] The mean of items contributing to each domain is taken as a domain score, and five domain scores are calculated for each individual child. To determine whether a domain score is in the *experiencing great difficulty*, *experiencing difficulty*, or *developing appropriately* category, all scores are ranked and compared against established national percentile boundaries and domain cut-offs. For the province of Alberta, the EDI results are reported as the number and percentage of children who fall into these three derived categories.[2] Having said that, the most widely used derived measure is vulnerability; if a child's score falls below the lowest 10th percentile in one or more of the five domains, a score of 1 (vulnerable) is given, otherwise, a score of 0 is given (not vulnerable). To put it differently, those classified as vulnerable are children who score low (below the 10th percentile cut-off of a comparison population, province or nation) in one or more of the five domains (Centre for Community Child Health and Telethon Institute for Child Health Research, 2009; Janus & Duku, 2007).

Analytical Methods

The analysis of MI is generally implemented in two frameworks: item response theory (IRT) and multi-group confirmatory factor analysis (MGCFA). The notion of MI in IRT has been referred to as DIF. In this study, we utilized MGCFA, a popular method to assess MI (see, Meredith, 1993) of EDI by means of Mplus 7.1 (Muthén & Muthén, 2012a). Since the items are all treated as categorical, the Weighted Least Squares Means and Variance adjusted (WLSMV) were chosen as estimators (Muthén & Muthén, 2012b). The Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI) and the Root Mean Square Error of Approximation (RMSEA) were used for assessing goodness of fit. The analytical procedure consisted of two major tests:

- CFA was performed for each age group and gender, and also overall sample to check goodness of fit of the hypothetical measurement model of each domain, postulated by EDI developers; and

- MGCFA was performed to test for configural, metric and scalar invariance across groups and testing of significant differences between those three models. There are two ways for assessing the differences between the three aforementioned models: using the chi-square difference test (Yuan & Bentler, 2004) or difference in fit indices (Byrne & van de Vijver, 2010). Due to the well-known sensitivity of the chi-square test of model fit to large sample size (Bagozzi, 1977; Bentler & Bonett, 1980), the difference in fit indices were used to evaluate measurement invariance across groups. A change in CFI and TLI less than 0.01 suggests evidence of invariance (Byrne & van de Vijver, 2010).

Following the work by Duku, Janus and Brinkman (2014) using EDI outcomes, the MI testing procedure recommended by Muthén and Asparhouhov (2002) was adopted. In this approach, thresholds and factor loadings are set free across groups while factor means should be fixed at 0 and scale factors should be fixed at 1. This serves as a less restricted model. Next, equality constraints are imposed on thresholds and factor loadings across groups while the factor means are fixed at 0, and scale factors are fixed at 1 in one group. The factor means and scale factors are set free in other groups. This will serve as a restricted model. If measurement invariance holds true, then there should not be a significant difference between the restricted and less restricted models. The primary software utilized is Mplus and the syntax that corresponds to the communication skills and general knowledge domain (i.e. the domain with smallest number of items) is presented in the Appendix. The procedures were administered for each domain. Additionally, descriptive statistics and reliability coefficients based on Cronbach's alpha were reported for each group as well as the overall sample.

Results

Results are presented in two parts as descriptive and MGCFA analysis for ESL and gender separately.

Descriptives

Descriptive statistics and domain scores across age groups and gender are presented in Tables 1 and 2.

As can be seen from Table 1, the majority of children in our sample are in the age range of 5 years and 3 months and above. Only 9.7% of children are 5 years and 2 months and below. And, the sample is more or less balanced in terms of gender ratio. In contrast, the majority of children in our sample are from families whose first language is not English.

Information in Table 2 is not surprising. As expected, the mean scores of all five developmental areas increase as the age of children increases. Further, standard deviations of developmental areas decrease while the age of children increases. This may mean that teachers' ratings are more consistent for older children. This could also mean that measuring developmental outcomes can become more objective if children are older. Girls were scored higher than boys across all domains, and children with English as their first language were also scored higher. Reliability coefficients across groups and overall sample are shown in Table 3.

Generally speaking, reliability coefficients are comparable across groups and are more or less close to these values for the overall sample. For physical health and well-being and language and cognitive development domains, although the differences are smaller in magnitude, the estimates are higher for children aged 5 years and 2 months and below, compared to other age

Table 1

Frequency distribution by groups

	Frequency	Percent	Cumulative Percent
5 years & 2 months and below	1,542	9.7	9.7
5 years & 3 months to 5 years & 6 months	4,738	29.8	39.4
5 years & 7 months to 5 years & 10 months	5,169	32.5	71.9
5 years & 11 months and above	4,472	28.1	100.0
Female	7,834	49.2	49.2
Male	8,087	50.8	100
ESL= Yes	11,823	74.3	74.4
ESL= No	4,068	25.6	100.0
Total	15,921	100.0	

Table 2

Descriptive Statistics of Five Developmental Areas by Age Groups and Gender

	G1		G2		G3		G4		Male		Female		ESL=Yes		ESL=No	
	M	St.D	M	St.D	M	St.D	M	St.D	M	St.D	M	St.D	M	St.D	M	St.D
PHYS	8.1	1.7	8.5	1.5	8.7	1.4	8.9	1.4	8.5	1.6	8.8	1.4	8.6	1.5	8.7	1.5
SOC	7.8	2.0	8.2	1.9	8.5	1.8	8.6	1.8	8.0	2.0	8.7	1.6	8.1	2.0	8.5	1.8
EMOT	7.8	1.6	8.0	1.5	8.2	1.5	8.3	1.5	7.7	1.6	8.5	1.3	7.9	1.5	8.2	1.5
LANG	7.6	2.2	8.1	1.9	8.4	1.8	8.7	1.6	8.1	1.9	8.5	1.7	7.8	2.1	8.5	1.7
COMM	6.4	3.0	7.1	2.8	7.6	2.7	8.0	2.6	7.1	2.9	7.8	2.7	5.7	3.0	8.1	2.4
Age	5.1	0.1	5.4	0.0	5.7	0.0	6.0	0.2	5.6	0.3	5.6	0.3	5.6	0.3	5.6	0.3

Note. G1= 5 years & 2 months and below; G2= 5 years & 3 months to 5 years & 6 months; G3= 5 years & 7 months to 5 years & 10 months; G4= 5 years & 11 months and above, St. D= Standard Deviation, M= Mean.

Table 3

Reliability Coefficient (Cronbach's Alpha)

	G1	G2	G3	G4	Male	Female	ESL=Yes	ESL=No	Overall
PHYS	0.802	0.788	0.793	0.785	0.801	0.783	0.797	0.793	0.795
SOC	0.959	0.957	0.958	0.958	0.959	0.955	0.959	0.958	0.958
EMOT	0.919	0.917	0.923	0.923	0.923	0.909	0.916	0.923	0.922
LANG	0.916	0.907	0.901	0.892	0.909	0.901	0.913	0.900	0.906
COMM	0.941	0.940	0.943	0.941	0.941	0.943	0.941	0.931	0.943

Note. G1= 5 years & 2 months and below; G2= 5 years & 3 months to 5 years & 6 months; G3= 5 years & 7 months to 5 years & 10 months; G4= 5 years & 11 months and above

groups and overall age. All estimates for boys show slightly higher reliability except for the communication skills and general knowledge domain where there seems to be higher internal consistency among items for girls. Furthermore, reliability estimates are also slightly higher for children with English as their first language except for the emotional maturity domain in which data from ESL children shows higher internal consistency.

CFA Analysis

The first step in testing MI is to check whether overall sample data fits the measurement model or not. This step was carried out using confirmatory factor analysis. For assessing goodness of fit in CFA models, Hu and Bentler (1999) suggested CFI and TLI values are equal to or greater than 0.95, and RMSEA values are equal to or smaller than 0.06. In addition, Byrne and van de Vijver (2010) suggested that RMSEA values as high as 0.08 represent a reasonable amount of approximation in the population.

The initial CFA analysis for the Physical health and well-being domain showed acceptable CFI and TLI values but an RMSEA value above 0.08. Following Duku, Janus and Birnkman (2014), we also allowed for cross-loadings in order to improve the fit without over-fitting the model. By allowing cross-loadings between the "*level of energy throughout the school day*" (Qa12) and "*overall physical development*" (Qa13) items (i.e., gross and fine motor skills or PHYS1) and the physical readiness for school day (PHYS2) sub-domain, acceptable fit was achieved. A cross-loading simply means that an item is related to more than one factor resulting in correlation between the factors because now they share common item(s). The cross-loading idea is theoretically an appropriate step when an item is believed to be conceptually related to any of the other factor(s). Figure 2 depicts the cross loading between PHYS1 and PHYS2 with dashed arrows.

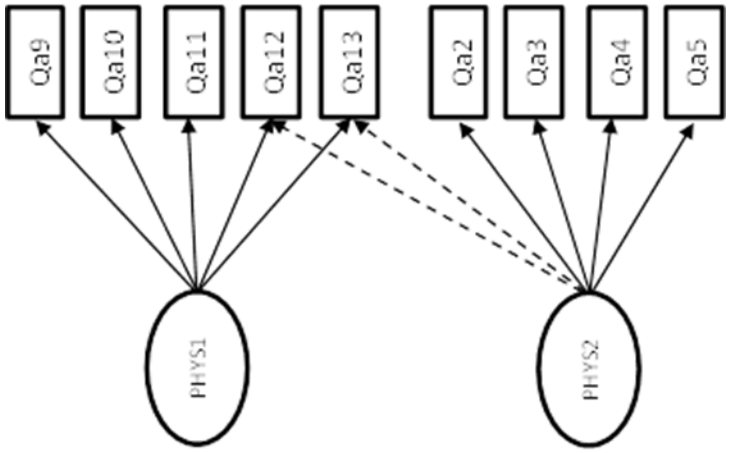

Figure 2. Cross-loading between PHYS1 and PHYS2

The Emotional maturity and Language and cognitive development domains showed acceptable fit and, therefore, no further modification was applied. As for the Social competence domain, we allowed cross-loading between "*shows self-confidence*" item (i.e., overall social competence) and readiness to explore new things sub-domain in addition to cross-loadings between "*takes care of school materials*" (i.e., responsibility and respect) and "*is curious about the world*" items (i.e., readiness to explore new things) and the subdomain, approaches to learning. Although Communication and general knowledge showed high CFI and TLI values, RMSEA was 0.104. Since this domain does not have any sub-domains, the only possible modification for this domain was to add error covariances between items, and we decided not to proceed with this domain. The main reason for not adding error covariances between items for the Communication and general knowledge domain was the fact that adding error covariances artificially accounts for unexplained variance without providing a rationale for doing so. Additionally, adding error covariances means that there are other factors outside of the model that can explain the variance-covariance of items, but we have no idea what those factors are. In such situations, it is better to alter the measurement model based on theoretical consideration than simply adding error covariance between items. Final results from CFA on each developmental area across overall sample are presented in Table 4.

Overall, there is acceptable fit in terms of CFI/TLI except for Communication and general knowledge where RMSEA value is above 0.1. It should be noted that CFI and TLI are comparative fit measures meaning that these indices assess the fit between estimated model and baseline model (i.e. a model in which factor loadings are fixed to one, error variances are fixed to zero and indicators assumed to be uncorrelated). In contrast, RMSEA is an exact fit measure that evaluates the approximate fit of the model to the population variance-covariance matrix (Schermelleh-Engel, Moosbrugger, & Müller, 2003). Large values for CFI/TLI show an acceptable improvement of model fit for almost all domains by using hypothetical models. However, RMSEA value suggests that the specified measurement model for the Communication and general knowledge domain might be mis-specified.

Measurement Invariance Across Gender

The MI analysis was administered only for domains that showed acceptable fit with the data. Results are presented in Table 5. As a reminder, Model0 represents the less restricted model and Model1 represents the restricted model where item thresholds and factor loadings were constrained to be equal across groups.

With respect to information in Table 5, all domains showed evidence of equality of item

Table 4

Results of CFA on Developmental Areas

	Chi-square	df	CFI	TLI	RMSEA
Physical health and well-being	2612.88	59	0.995	0.993	0.052
Social competence	22462.49	291	0.976	0.973	0.069
Emotional maturity	23755.63	401	0.976	0.974	0.060
Language and cognitive development	11323.45	295	0.969	0.966	0.046
Communication and general knowledge	3443.73	20	0.996	0.994	0.104

thresholds and factor loadings across gender. All domains showed fit improvement in terms of CFI, TLI and RMSEA after imposing model constraints. The improvement in the model fit was minimal for Language and cognitive development compared to other domains. The major improvement occurred for the Emotional maturity domain in terms of TLI and for the Physical health and well-being domain in terms of RMSEA.

Measurement Invariance Across ESL Status

Results of the MI analysis across ESL status are presented in Table 6.

Based on the information in Table 6, all domains showed evidence of equality of item thresholds and factor loadings across ESL status. All domains showed fit improvement in terms of CFI, TLI and RMSEA after imposing model constraints. The improvement in the model fit was minimal for Language and cognitive development compared to other domains. The major improvement occurred for Physical health and well-being domain in terms of RMSEA.

The RMSEA for Physical health and well-being domain showed the largest difference between the models for both ESL status and gender. This may suggest that regardless of high values of CFI and TLI, the measurement model suffers from some level of mis-specifications. A

Table 5

MI of EDI Domains Across Gender

		CFI	TLI	RMSEA	ΔCFI	ΔTLI	ΔRMSEA
Physical health and well-being	Model0	0.991	0.988	0.065			
	Model1	0.993	0.992	0.055	0.002	0.004	-0.010
Social competence	Model0	0.976	0.974	0.066			
	Model1	0.980	0.979	0.059	0.004	0.005	-0.007
Emotional maturity	Model0	0.973	0.970	0.062			
	Model1	0.976	0.976	0.056	0.003	0.006	-0.006
Language and cognitive development	Model0	0.971	0.968	0.047			
	Model1	0.971	0.969	0.046	0.000	0.001	-0.001

Table 6

MI of EDI Domains Across ESL

		CFI	TLI	RMSEA	ΔCFI	ΔTLI	ΔRMSEA
Physical health and well-being	Model0	0.991	0.988	0.067			
	Model1	0.993	0.992	0.055	0.002	0.004	-0.012
Social competence	Model0	0.978	0.976	0.066			
	Model1	0.981	0.981	0.059	0.003	0.005	-0.007
Emotional maturity	Model0	0.975	0.973	0.061			
	Model1	0.978	0.978	0.055	0.003	0.005	-0.006
Language and cognitive development	Model0	0.969	0.966	0.048			
	Model1	0.969	0.967	0.047	0.000	0.001	-0.001

further look at this issue revealed that adding error covariance between "*proficiency at holding a pen, crayons, or a brush*" and "*ability to manipulate objects*" items from gross and fine motor skills sub-domain can rectify this issue.

Comparison Across Gender and ESL Status

Table 7 presents mean scores of EDI domains across gender and ESL status. The last column includes effect size measures calculated based on the Cohen's d for the two preceding means.

Results in Table 7 suggest that for domains with measurement invariance across gender and ESL status, the observed difference is not considerable as indicated by effect size values. All Cohen's d values show small effect sizes except Language and cognitive development, which has a value close to medium effect size across gender (i.e., 0.41 for female and 0.35 for male). The major concern is the Communication and general knowledge domain that shows a large effect size (shown in bold face font), as evidenced by the observed means. This is the model where we did not achieve acceptable fit in the overall sample. This might be due to heterogeneity of data. A further look into this domain revealed that adding error covariances between "*ability to listen in English*" and "*ability to understand on first try what is being said to him/her*" items as well as "*ability to communicate own needs in a way understandable to adults and peers*" and "*ability to articulate clearly, without sound substitutions*" items can reduce RMSEA from 0.104 to 0.078.

Discussion and Conclusion

The EDI is arguably the most widespread of early development screening tools and is continuing to expand its use. For example, an Australian adaptation of the Canadian EDI —the Australian Early Development Index (AEDI) —has now been used in over 60 communities across Australia (Halfon, Russ, Oberklaid, Bertrand, & Eisenstadt, 2009). A common concern with any standardized test is construct bias, for instance, whether the tool measures what it purports to measure equivalently across groups. Generally speaking, a test is construct-biased if the

Table 7

MI of EDI Domains Across ESL

		ESL= No	ESL= Yes	Cohen's d
Female	Physical health and well-being	8.87	8.74	0.10
	Social competence	8.81	8.46	0.22
	Emotional maturity	8.52	8.30	0.17
	Language and cognitive development	8.65	7.94	0.41
	Communication and general knowledge	8.36	5.96	**0.98**
Male	Physical health and well-being	8.47	8.39	0.05
	Social competence	8.09	7.74	0.18
	Emotional maturity	7.79	7.59	0.12
	Language and cognitive development	8.28	7.62	0.35
	Communication and general knowledge	7.76	5.39	**0.89**

outcomes tend to be more biased to one group than another, suggesting that the test items in fact function differently across groups. That is, if developmental outcomes vary radically between two age cohorts, such as 4 and 6 year olds or two cultures, such as Spanish and Turkish, and if the two were treated as one, they are likely to lose construct validity (Cook & Campbell, 1979). Even if the tool itself is not demographically or culturally biased, the contexts in which the tool is administered may lead to variations in outcomes due to a number of reasons, such as family and community characteristics, genetics, neuroscience and early brain development, health of children, and parenting skills, to name a few (Duku, Janus & Brinkman, 2014; Rock & Stenner, 2005). Consequently, we assume that there is measurement non-invariance, that is, the tool does not measure the outcomes in the same manner for different groups. Unfortunately, to the authors' knowledge, test and item biases of EDI are given very little attention (e.g., Hymel, LeMare & McKee, 2011; Janus, Brinkman & Duku, 2011), even though there can be possible sources of bias.

In this study, we examined measurement invariance of EDI domains across ESL status and gender by means of multi-group confirmatory factor analysis. We first analyzed goodness of fit of measurement model in the overall sample in order to have an idea on how well models fit the data. With minor modification for two domains, results of separate CFA showed acceptable goodness of fit in terms of CFI/TLI but non-acceptable fit for Communication and general knowledge domain in terms of RMSEA. This means that the specified model does not fit population variance-covariance reasonably. Thus this domain was excluded from MI analysis.

The measurement invariance analysis revealed that there is acceptable evidence of measurement invariance across groups for all four domains. This finding implies that domain scores have the same meaning and metric across ESL status and gender. The only concern that arose with MI analysis was the considerable reduction in RMSEA for the Physical health and well-being domain in the restricted model. This implies that the measurement model for this domain has potential for improvement. A supplementary analysis for Physical health and well-being and Communication and general knowledge domains indicated that adding error covariances can improve the model fit in terms of RMSEA. An error covariance between two items can be interpreted as one or more additional factors affecting these items that are not considered in the current measurement model. Another possible likely interpretation could be overlap between the two items. In other words, the two items are perceived in the same way by respondents, thereby adding redundancy to the model.

As Borsboom (2006) and Steinmetz (2013) had mentioned, MI is critical since we deal with latent variables that cannot be measured directly, and any indirect measurement is prone to error. This error or bias is not only important in the use of composite observed scores for further analyses (within ANOVA), but it also matters when the instrument is going to be used for selecting or classifying examinees (children in this case). As for the EDI, domain scores are used for classifying children as vulnerable or not and developing appropriately or not. Therefore, any bias in the measurement model such as measurement non-invariance across groups should be taken seriously. In practice, when group comparisons are made, summed scores of items within each domain are used. Therefore, achieving MI is a requirement for group comparisons because we cannot compare groups on different things with different scales.

This study has strengths. To our knowledge, the most important is that no study has examined measurement invariance of the original form of EDI outcomes across ESL status and gender within the framework of multi-group confirmatory factor analysis. Moreover, this study provides some validation basis not only for previous studies by researchers like Guhn et al.,

(2007) but also for future studies that might use domain score for analysis. Nevertheless, this study showed that any group comparison for Communication and general knowledge domain should be interpreted with caution. There are several limitations as well with the present study. We only analyzed data with respect to the original measurement model, postulated by the EDI's own developers. This means we did not consider any major model modifications in terms of omitting/changing items. Further, the present study did not attempt to explore MI models in terms of auxiliary factors, such as language. Consideration of relevant factors (e.g., socioeconomic background) could give more insight into the causes of variation and ultimately improve the reliability and validity of the tool. Despite these limitations, our results highlight the complexities of using EDI data for all preschool children, ranging from ages 4 to 7 years with the assumption that one size fits all, if we really want to do justice to the issue of early child development outcomes and child development in general.

References

Bagozzi, R. P. (1977). Structural equation models in experimental research. *Journal of Marketing Research, 14*, 209-226. doi:10.2307/3150471.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606. doi:10.1037/0033-2909.88.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44*(11), S176-S181.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in largescale cross-cultural studies: addressing the issue of nonequivalence. *International Journal of Testing, 10*(2), 107–132.

Centre for Community Child Health and Telethon Institute for Child Health Research (2009). *A snapshot of Early Childhood Development in Australia–AEDI National Report*, Australian Government.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Commonwealth of Australia, (2013). *Australian Early Development Index 2012: Summary report* (updated November 2013). Canberra: Department of Education.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95*, 134–135.

Duku, E., Janus, M., & Brinkman, S. (2014). Investigation of the Cross-National Equivalence of a Measurement of Early Child Development. *Child Indicators Research*, 1-19.

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*(3), 453-472.

Halfon, N., Russ, S., Oberklaid, F., Bertrand, J., & Eisenstadt, N. (2009). An International Comparison of Early Childhood Initiatives: From Services to Systems. Commonwealth Fund pub. No. 1241 Retrieved from http://www.commonwealthfund.org/Publications/Fund-Reports/2009/May/An-international-Compar.

Hymel, S., Le Mare, L., & McKee, W. (2011). The Early Development Instrument (EDI): An examination of convergent and discriminant validity. *Social Indicators Research, 103*(2), 267-282.

Janus, M., & Duku, E. (2005). *Development of the short Early Development Instrument (S-EDI)*. Report for the World Bank. Retrieved from https://edi.offordcentre.com/wp/wp-content/uploads/2015/06/REPORT.short_edi_june2005.pdf

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*(3), 375-403.

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., Brinkman, S. A., & Duku, E. K. (2011). Validity and psychometric properties of the Early Development Instrument in Canada, Australia, United States, and Jamaica. *Social Indicators Research, 103*(2), 283-297.

Krishnan, V. (2013). *The Early Child Development Instrument (EDI): An item analysis using Classical Test Theory (CTT) on Alberta's data*. Retrieved from http://www.cup.ualberta.ca/wp-content/uploads/2013/04/ItemAnalysisCTTCUPWebsite_10April13.pdf.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Muthén, B.,& Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Mplus Web Notes: No. 4. Retrieved from www.statmodel.com

Muthén, L. K., & Muthén, B. O. (2012a). Mplus 7.1 for Windows. Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2012b). Mplus User's Guide (7th Ed.). Los Angeles: Muthén & Muthén.

Rock, D. A., & Stenner, A. J. (2005). Assessment issues in the testing of children at school entry. *The Future of Children, 15*(1), 15-34.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Schmitt, N., Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 4, 210-222.

Steinmetz, H. (2013). Analyzing Observed Composite Differences Across Groups. Methodology: *European Journal of Research Methods for the Behavioral and Social Sciences, 9*(1), 1-12.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486-492.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Yuan, K.-H., & Bentler, P.M. (2004). On chi-square difference and z-tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*, 737-757.

Notes

1 Changes were made to the original version of the instrument, which included measures especially of five-point scales in both physical health and well-being and communication skills & general knowledge areas (Janus & Offord, 2007).

2 One may find different categorizations as developmentally vulnerable, developmentally at risk, and on track in literatures and reports dealing with EDI (see, Commonwealth of Australia, 2013).

Amin Mousavi is an Assistant Professor of Psychometrics, Classroom Assessment, and Measurement at the Department of Educational Psychology and Special Education, College of Education, University of Saskatchewan. His main area of interest for teaching and research is psychometrics and quantitative methodology in Social and behavioral sciences.

Vijaya Krishnan is a Researcher at the School of Public Health at University of Alberta, Edmonton. She is specialized in technical demography and has written and published widely in population health and early child development. Vijaya received her Masters' in Mathematics and Masters' in Demography from the University of Kerala. She received her doctorate in Demography from University of Alberta in 1989. Since then, she held various positions in government, academia, and private sector, which includes Alberta Health, Simon Fraser University, and University of Botswana. She is the recipient of several research grants and scholarship.

Appendix

Sample Mplus syntax for assessing MI of the Communication and general knowledge domain across ESL groups. Note that all the texts after the exclamation mark (!) are considered as comments in Mplus (in green).

```
TITLE: Measurement Invariance of EDI domains-less restricted model;
Data:
 file is "data-mplus.csv"; ! Calling data

Variable:
Names=Qb1 Qb2 Qb3 Qb4 Qb5 Qb6 Qb7 Qc26 gender EFSL; ! Variable names
Missing are all(99); ! Define the code for missing values
Usevariables =Qb1 Qb2 Qb3 Qb4 Qb5 Qb6 Qb7 Qc26 EFSL; ! List of variables to be used in
analysis
Categorical=Qb1 Qb2 Qb3 Qb4 Qb5 Qb6 Qb7 Qc26; ! Name of categorical variables
Grouping is EFSL (0=no 1=yes); ! Define grouping variable and group labels

Analysis:
 TYPE= general;
 Estimator= WLSMV;

 MODEL:
 COMM BY Qb1@1 Qb2 Qb3 Qb4 Qb5 Qb6 Qb7 Qc26; ! Define measurement model

 [COMM @0]; ! Fix factor mean at 0 for both groups
 {Qb1-Qc26@1}; ! Fix scale factors at 1 for both groups

 MODEL yes: ! Define group-specific model, the ESL group in this example
 COMM BY Qb1@1 Qb2 Qb3 Qb4 Qb5 Qb6 Qb7 Qc26; ! Estimate factor loadings for this group
 [Qb2$1-Qc26$1]; ! Estimate item intercepts for this group
 [Qb2$2-Qc26$2]; ! Estimate item intercepts for this group
```

Now after running the above syntax, the next step is to modify just the model specifications for the group with ESL status (i.e., ESL = yes) as follows:

```
TITLE: Measurement Invariance of EDI domains- restricted model;
Data:
 file is "data-mplus.csv"; ! Calling data

Variable:
Names=Qb1 Qb2 Qb3 Qb4 Qb5 Qb6 Qb7 Qc26 gender EFSL; ! Variable names
Missing are all(99); ! Define the code for missing values
Usevariables =Qb1 Qb2 Qb3 Qb4 Qb5 Qb6 Qb7 Qc26 EFSL; ! List of variables to be used in
analysis
Categorical=Qb1 Qb2 Qb3 Qb4 Qb5 Qb6 Qb7 Qc26; ! Name of categorical variables
```

```
Grouping is EFSL (0=no 1=yes); ! Define grouping variable and group labels

Analysis:
 TYPE= general;
 Estimator= WLSMV;

 MODEL:
 COMM BY Qb1@1 Qb2 Qb3 Qb4 Qb5 Qb6 Qb7 Qc26; ! Define measurement model

 [COMM @0]; ! Fix factor mean at 0 for both groups
 {Qb1-Qc26@1}; ! Fix scale factors at 1 for both groups

 MODEL yes: ! Define group-specific model, the ESL group in this example
 [COMM]; ! Estimate factor mean for this group
 {Qb1-Qc26};; ! Estimate scale factors for this group
```