Investigating Key Psychometric Properties of the French Version of the Early Years Evaluation-Teacher Assessment

Robert Laurie
University of New Brunswick

Elizabeth Sloat
University of New Brunswick

Abstract

This research investigates key psychometric properties of the French Early Years Evaluation-Teacher Assessment measure designed to systematically assess kindergarten children across five social and academic developmental domains: awareness of self and environment, social skills and behaviour, cognitive abilities, language and communication, and physical development. New Brunswick francophone kindergarten children were recruited to assess the instrument's internal consistency; content, construct, concurrent and discriminant validity; and linguistic bias relative to the English version. Results indicate that the French measure has strong psychometric properties, and that it can therefore be used with confidence to screen for at-risk children in francophone kindergartens.

Canadian Journal of Education / Revue canadienne de l'éducation 39:3 (2016)

www.cje-rce.ca

Keywords: child development, Early Years Evaluation, screening test, kindergarten, school readiness, vulnerable children, reliability, validity, linguistic bias, differential item functioning (DIF)

Résumé

Cette recherche étudie des propriétés psychométriques de la version française de l'outil Early Years Evaluation-Teacher Assessment. L'outil sert à mesurer le développement des enfants à la maternelle dans cinq domaines essentiels pour la réussite scolaire et sociale: conscience de soi et de l'environnement, habiletés sociales et comportement, habiletés cognitives, langue et communication, et développement physique. Des enfants francophones de la maternelle au Nouveau-Brunswick ont été recruté pour évaluer la fidélité de l'instrument; sa validité de contenu, de construit, et sa validité concourante et discriminante, ainsi que son biais linguistique par rapport à la version anglaise. Les résultats indiquent que les propriétés psychométriques de l'outil de mesure sont excellentes et qu'il peut être utilisé en toute confiance pour dépister les enfants francophones à risque dès la maternelle.

Mots-clés : développement de l'enfant, Évaluation de la petite enfance, test de dépistage, maternelle, préparation à l'école, enfants vulnérables, fidélité, validité, biais linguistique, fonctionnement différentiel des items (FDI)

Introduction

Adapting seamlessly to the school environment at entry to kindergarten and achieving success in the first years of school is highly dependent on children's abilities, behaviours, and attitudes (Canadian Council on Learning–Conseil Canadien de l'Apprentissage [CCL-CCA], 2006). Research shows that differences in these traits during the early years can predict later school achievement (Cabell, Justice, Konold, & McGinty, 2011; McCartney, 2007). Children entering kindergarten already behind their peers typically fall further behind each passing year unless they receive early and targeted intervention and support. It is possible to alter children's poor growth trajectories, but identifying clearly and accurately where each individual child struggles along the continua of multiple developmental domains is central to providing the direct instruction needed to address achievement differences (Canadian Education Statistics Council [CESC], 2009). Early screening benefits all children, but vulnerable children in particular must have needs addressed at the earliest age possible if they are to have the best chance of overcoming difficulties (Doherty, 1997; Fox, Dunlap, & Cushing, 2002; Lyon et al., 2001).

Many education jurisdictions in Canada and internationally now collect diagnostic information on kindergarten children's development using standardized measures of assessment (e.g., Daily, Burkhauser, & Halle, 2010). Systematic and ongoing data collection and progress monitoring strategies are implemented at school entry so that social and cognitive issues, reading delays and any other problem areas are identified early in a child's developmental trajectory. The move to early and regular progress monitoring across multiple domains has marked a major shift in educational practice in recent years from the traditional wait-to-fail approach (Greenwood, Bradfield, Kaminski, Linas, Carta, & Nylander, 2011; Sloat, Beswick, & Willms, 2007). Rather than waiting for children to present with clearly established learning disabilities over subsequent years of schooling, the approach now is one of prevention, which relies on early problem identification and targeted intervention, so learning challenges can be corrected before they reach disabilities status (Greenwood et al., 2011).

In Canada, the Early Years Evaluation-Teacher Assessment (EYE-TA; The Learning Bar, 2016) is used in every province to screen kindergarten children for potential delays in five developmental domains foundational to early learning and overall success: (1) Awareness of Self and the Environment, (2) Social Skills and Behaviour, (3) Cognitive Abilities,

(4) Language and Communication and (5) Physical Development. As a standardized assessment, the measure is effective because it provides a systematic framework for informing teachers' and administrators' decisions about the early learning and support needs of each child.

The EYE-TA, however, like many early childhood assessment instruments, is in English, and serves only English populations. In many provinces like Alberta, Ontario, Quebec, and New Brunswick, there are large francophone populations who need, and should be able, to benefit equally from early screening and instructional intervention and support, and yet few effective French standardized assessment measures exist (Thordardottir, Keheyia, Lessard, Sutton, & Trudeau, 2010). Some provinces (New Brunswick, Prince Edward Island, Newfoundland and Labrador, Saskatchewan, and British Columbia) now use a French version of the EYE-TA, the Évaluation de la petite enfance–appréciation de l'enseignante (ÉPE-AE), to screen francophone kindergarten students. While the English EYE-TA has strong reliability and validity psychometric properties (KSI Research International, 2009), similar information is not known about the ÉPE-AE. This study fills this knowledge gap by investigating key psychometric properties of the ÉPE-AE. Three questions therefore guided our work: (1) Is the ÉPE-AE a reliable measure for assessing children's developmental status? (2) To what extent does the ÉPE-AE show strong content, construct, and convergent and divergent validity? (3) Does the ÉPE-AE or the EYE-TA show bias in favour of one language group over the other?

The EYE-TA and ÉPE-AE Assessment Measures

We turn here to a description of the EYE-TA and its French version, the ÉPE-AE, to provide an overview of how the measure is designed and administered, and how feedback is reported to educators and schools for the purpose of informing early instructional interventions and support. We noted above the five distinct but connected domains of emergent literacy, readiness for school, and academic success included in the measures, as first suggested by the National Education Goals Panel (NEGP) in 1991, and later endorsed by the National Research Council's (NRC) committee report on developmental outcomes and assessments for young children: (1) physical well-being and motor development,

(2) social and emotional development, (3) approaches toward learning, (4) language usage, and (5) cognition and general knowledge (NRC, 2008; see also CCL-CCA, 2008; Doherty, 1997; National Governance Task Force on School Readiness, 2005; National School Readiness Indicators Initiative, 2005; Stedron & Berger, 2010). These same domains are included in the EYE-TA and the ÉPE-AE assessment measures that comprise a systematic framework for measuring a kindergarten child's development.

The Awareness of Self and Environment domain assesses a child on aspects of general knowledge and understanding; for instance, the role of community members like police and doctors, and on relational concepts such as front-and-back, and first-and-last. Social Skills and Behaviour assesses children's social and behavioural interactions in the school setting to provide an indication of how children approach new learning situations, their ability to adhere to classroom rules, and whether they exhibit signs of hyperactivity, inattention, anxiety, emotional difficulties, or physical aggression. Cognitive Ability includes a set of items for assessing mathematics, problem-solving, and pre-reading skills including number counting, phonological awareness, and letter recognition. Language and Communication assesses both receptive and expressive oral language capabilities and includes items directly related to communicative functioning in the classroom. Finally, the fifth domain, Physical Development, assesses fine and gross motor development, from hand-eye coordination to the physical coordination necessary for playing with other children.

The lists of questions, or assessment scales, teachers complete relative to each developmental area range from seven items in the Awareness of Self and Environment and Language and Communication domains, to as many as 15 items in the Social Skills and Behaviour domain. The four response categories for all but the Social Skills and Behaviour domain are simply worded while also prompting teachers to make clear, concrete judgements about performance on each scale item. A score of one indicates that a child is "unable to do it," while a score of two indicates that a child "can do it partially." A three score indicates a child "can usually do it," while a score of four means that children "can do it consistently." Since the Social Skill and Behaviour domain screens for potential social, emotional, and behavioural challenges, ratings for items in this scale target the frequency with which particular behaviours are evident such that responses range from a score of one, "regularly (nearly every day)"; two, "occasionally (about once a week)"; three, "once in a while (about once a month)"; to four, "never or rarely."

Several features of the EYE-TA and ÉPE-AE make these measures unique to other assessment tools. The assessment is conducted with all children and not just a select or representative few, since individualized results are by far the better option when the objective is to identify and target every child's learning needs as part of a comprehensive and longitudinal development monitoring system. The EYE-TA and ÉPE-AE also differ from other kindergarten school readiness instruments (e.g., Canadian Psychology Association, 1995; Janus & Offord, 2007) in that it is a skills-based assessment requiring children to demonstrate actively their knowledge and skills. Performance-based measures like the EYE-TA and ÉPE-AE require educators to know for certain whether each child is able or unable to complete a specific skill or task. To facilitate ensuring teachers are able to make precise assessments, pictures and other support tools are available in both English and French from the instrument's website for conducting quick direct assessments so ability is determined accurately rather than based on perceptions alone about a child's knowledge and abilities.

When completing the assessment, teachers are urged to rate all children on each item at the same time, rather than rating individual children on all items across all five domains at once. Following this approach is important because it fosters consistency in evaluator expectations for performance and ability levels as each criterion is assessed. This mean that teachers would, for example, assess all children on the number counting item in the Cognitive Ability domain before moving on to complete another assessment item for all children, such as alphabet recognition. Assessing in this manner aids in minimizing the halo effect we can often derive, even unknowingly, from factors like socio-economic status or likeable personalities, which can influence our perceptions about a child's knowledge, skill, and ability (Thorndike, 1920). Finally, conducting the assessment and collecting data for each child is relatively easy for teachers to accommodate within normal teaching schedules and school day timeframes. Rather than pulling children out of regular instruction, as is often necessary when required to assess directly a child's knowledge or ability, the EYE-TA and ÉPE-AE are completed over a few weeks during regular classroom instruction so teachers have time and opportunity to observe and assess all children accurately on all domain items.

As assessments are completed, teachers are provided login details for entering their individual ratings for all scale items for each child through a secure website. A status report setting out a class list of results showing each child's performance on each of the

five domains is promptly generated once all scores have been entered. The reports are simply designed so educators and administrators can easily read and interpret a child's results. A green-coloured marker next to a domain indicates that performance is "at appropriate development," while a yellow-coloured marker indicates that a child is "experiencing some difficulty." A red-coloured marker indicates that there is "evidence of significant difficulty." Schools then have the information they need to provide continued high-quality classroom instruction in combination with a secondary level of targeted support for those experiencing some difficulty, and tertiary intensive intervention for children experiencing significant difficulty. Continued diagnostic assessments using more domain-specific and detailed instruments like the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2003) and the Phonological Awareness Literacy Screener (PALS; Invernizzi, Sullivan, Meier, & Swank, 2004) to assess emergent literacy knowledge, are equally essential in providing a comprehensive monitoring system to track kindergarten children's early and continuous knowledge and skill.

Given the merits of the ÉPE-AE for francophone populations as a universal screener on multiple domains of early learning and development, examining key psychometric properties of the measure is important. The results of our study ensure that francophone children equally have access to, and can benefit from, a well-designed and comprehensive early childhood development screening measure. We now set out our methodological procedures and study results in the following sections.

Method

EYE-TA Translation

Our first step was to obtain highly accurate translations for the EYE-TA, its scoring rubric, administration instructions, and the web-based support tools available for directly assessing children as needed. All materials were translated to French by a professional translator. Documents were then independently checked and verified by two staff members of the New Brunswick Department of Education and Early Childhood Development based on their bilingual expertise and background working in French and large-scale assessment contexts. Reviewers deemed translations appropriate, word counts were

generally the same in both languages, and administration procedures and item formats were similar. A final step provided participating francophone kindergarten teachers the opportunity to offer any minor modifications they thought might facilitate the ÉPE-AE's administration. This step helped both to clarify the administration procedures while ensuring a common and appropriate understanding and interpretation of each item and the scoring rubric while maintaining the integrity of the French form in relation to the English version.

Participants and Setting

Twelve schools in a francophone school district covering the southern portion of New Brunswick, as one of the provinces using the ÉPE-AE, were approached to participate in our study. This convenience sample met the need to assess francophone students from comparable socio-economic areas and access to health and wellness community services similar to anglophone students given that the English and French school districts overlap geographically. Eleven female francophone kindergarten teachers from six schools volunteered for study participation, some of whom were recent BEd program graduates, while others had several years of teaching experience. A letter explaining the study's purpose was sent to all parents seeking permission to include their child's assessment data in the research. Letters were issued and returned through classroom teachers, which ensured a high parental response rate. Complete data sets for 193 francophone kindergarten students (48.8% boys and 51.2% girls), ranging between 5.1 and 6.3 years (mean = 5.7, *S.D.* = .29), were included in the analyses. A complete EYE-TA data set comprised of 389 anglophone kindergarten students (55.5% boys and 44.5% girls) aged 5.1 to 7.0 years (mean = 5.6, *S.D.* =.31) obtained in a previous study was made available for analytical and comparative purposes.

Teachers' Training

Teachers received a detailed training session to ensure a common and appropriate interpretation of each ÉPE-AE item and its completion either through observation during normal school activities or through assessing skills directly using downloadable support materials. Teachers were shown how to enter student data on a secure website following

an item-by-item rather than child-by-child procedure, in part toward reducing the potential for bias due to the halo effect (Thorndike, 1920).

Data Collection Procedures

Data collection occurred over a one-month period following the professional development session on the ÉPE-AE. In keeping with the administration format of the instrument, teachers had a one-month period in which to observe, and assess directly as needed, all students on all items of the five domains. They were provided login details for the secure website so they could access direct assessment support materials and enter their assessment responses for each child. During this same period, data were also collected on each child using two additional instruments for determining the ÉPE-AE's concurrent and discriminate validity. These two types of validity were included to see which ÉPE-AE domains showed strong correlations with other assessments designed to measure the same thing (concurrent validity), and which domains showed weaker correlations with other assessments designed to measure different things (discriminant validity). To this end, eight newly retired, highly experienced kindergarten teachers were recruited and trained to administer to all francophone children the French version of the Peabody Picture Vocabulary Test (PPVT), or the Échelle de vocabulaire en images Peabody (ÉVIP; Dunn, Dunn, Leota, Lloyd, & Thériault-Whalen, 1993), and the reading subtest of the French Canadian version of the Weschler Individual Achievement Test (WIAT–II; Wechsler, 2001). The ÉVIP is a standardized measure of a child's receptive vocabulary, while, at the kindergarten level, the WIAT-II subtest assesses emergent and early reading knowledge such as phonological awareness, decoding (letter naming), and word reading skills. Both the ÉVIP and the French WIAT–II were ideally suited for this study since they are considered "gold standards" in standardized literacy assessments. The PPVT is a standardized measure of verbal ability widely used since 1959, and the WIAT has also been used internationally since its release in 1992. As such, the results of these assessments are well suited for comparative purposes with the ÉPE-AE.

Data Analysis

To address the three research questions guiding our study, a number of statistical analyses were conducted to determine the instrument's internal consistency reliability, content

validity, construct validity, and concurrent and discriminant validity. Reliability is often synonymous with consistency, stability, and predictability (Hubley & Zumbo, 1996). As such, internal consistency determines whether an instrument yields consistent results when used in similar conditions, and the extent to which its items designed to measure the same construct produce similar results, even under differing assessment conditions (Zwyno, 2003). The measure is based on the correlations between all of the ÉPE-AE items, or those of its individual subscales within each of its five domains. Internal consistency reliability was calculated for each of the five ÉPE-AE domains and reported using Cronbach's coefficient alpha (α). An α score of .70 is considered satisfactory in a social science research investigation such as this one (Nunnally, 1978).

Determining an assessment's content validity is typically determined based on relying on the knowledge and expertise of those familiar with the constructs being measured, which for the EPE-AE are its five developmental domains. As such, content validity was addressed in the original design of the EYE-TA and reiterated in the ÉPE-AE based on wide agreement in the research, policy, and practice literatures governing the five domains to be monitored along with the assessment scale items used in each domain (NSRII, 2005; NRC, 2008; Stedron & Berger, 2010). Further, since participating teachers were asked to comment on and discuss the relevance of each item during their training to ensure that they found the ÉPE-AE suitable for assessing a child's early development, and that they understood and adhered to the purpose of each item, their analysis and comments added to the ÉPE-AE's content validity. Obtaining input from teachers is an approach also used by other researchers (Vasilyeva, Ludlow, Casey, & St. Onge, 2009) as it allows practitioners to complement theories with examples of their applications in classroom settings. In our study, teacher input not only clarified links between theory and practice but also helped to build a common understanding of these links. In turn, this common understanding served to increase the ÉPE-AE's reliability given the consistent interpretation by all teachers of its skills-based questions.

The ÉPE-AE's construct validity—the extent to which the items and scales used in an assessment actually provide information on the constructs they are designed to measure—was obtained separately for each domain by carrying out a principal components analysis (Garson, 2013). In preparing for this analysis, however, it was important to demonstrate that the data in each domain are suitable for a principal components analysis prior to carrying out the final construct validity analysis (Huck, 2008). To ensure that the

strength of the variables in each ÉPE-AE domain was sufficient to continue with factor analysis, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy coefficient was obtained for each developmental domain. The KMO coefficient should be greater than .60 to continue with the analysis (Tabachnik & Fidell, 2001). Since KMO results for all domains were .80 and above, factor analysis results were then interpreted using a Scree plot and their eigenvalues respecting the Kaiser-Guttman rule, which states that factors must have eigenvalues greater than 1.0 (Kaiser, 1960). KMO coefficients and factor analyses are reported for each domain in Table 1 below.

There is evidence of concurrent validity when measures, which theoretically should be related to each other, show strong correlations. Conversely, there is evidence of discriminant validity when measures, which theoretically should not be related to each other, show weak correlations. Concurrent validity and discriminant validity measures were obtained by comparing the results of a Pearson correlation between the ÉPE-AE data in each domain and the student scores on the Échelle de vocabulaire en images Peabody (EVIP) and the French Canadian Weschler Individual Achievement Test (WIAT–II) reading subtest. The coefficient of determination (the estimated strength of the relationship) was reported (see Table 2) since this allows for a more appropriate interpretation of the shared variance between variables.

Measurement bias—often called differential item functioning (DIF)—is present when respondents of equal ability or skill have a different probability of correctly answering the same question on a test or questionnaire. Extant data from 359 anglophone kindergarten students were used along with that of the 193 francophone kindergarten students who completed the ÉPE-AE to test for measurement bias based on language. Two different methods were used: item response theory (IRT), following Raju's method of determining DIF (Oshima & Morris, 2008); and the Mantel-Haenszel method (Mantel & Haenszel, 1959). Both tests were required to identify an item as being biased based on language for it to be considered a biased, or DIF, item. Items identified by only one of the two methods were not considered biased (see Table 3).

Results

Reliability (internal consistency). The reliability, or internal consistency, of the entire ÉPE-AE was .91. Domain Cronbach coefficient alpha values ranged from a low of

.77 to a high of .92. While most values were concentrated between .80 and .84, lower values were clustered in the Physical Development domain with higher values concentrated in Language and Communication. Values reporting results based on single item deletions are presented in Table 1.

Construct validity. The KMO coefficient for each of the five domains was well above the .60 threshold recommended for continuing with principal components analysis, the values of which are reported for each domain in Table 1. Coefficients ranged from .80 for the Physical Development domain, to a high of .90 for the Language and Communication domain. Scree plot and eigenvalue analyses indicated that the ÉPE-AE domains have between one and four factors. Finally, the percentage of the total variance explained by the factor(s) for each domain, as well as the associated factor loadings and extracted commonalities, are also reported in Table 1.

Table 1. Reliability and construct validity of the ÉPE-AE domains

Item	**ÉPE-AE Domains**		**Total variance explained (%)**	**Construct validity**				
	Awareness of Self and Environment (α= .83; KMO = .86)	**α if item deleted**	**50.8**	Factor loading	Factor loading	Factor loading	Factor loading	Extracted commonalities (h^2)
1	recognize unfamiliar animals	.80		.738				.544
2	understand time relative to daily routines	.80		.753				.567
3	identify community member's roles	.81		.686				.471
4	identify items belonging to the same category	.82		.603				.364
5	complete analogue sentences	.81		.693				.480
6	understand relational concepts	.80		.728				.531
7	describe the function of familiar objects	.80		.773				.597
	Social Skills and Behaviour (α = .85; KMO = .84)		**70.1**					
1	is sad or depressed	.85		.466	.503			.544
2	harms others physically	.83		.753		.478		.797
3	has difficulty staying seated	.84		.622	-.414			.621
4	follows directions	.84		.647				.701
5	seems scared or anxious	.85			.713			.721
6	kicks or hits peers	.83		.759				.750
7	has difficulty staying on task	.83		.718				.767
8	worries a lot	.85			.620			.654

9	persists in the face of adversity	.86					.771	.704
10	intimidates peers	.84		.735		.489		.785
11	seems nervous and tense	.84		.547	.584			.774
12	shows interest in class activities	.84		.624				.609
13	is mean and cruel toward peers	.84		.675		.581		.800
14	has difficulty staying attentive	.83		.672	-.452			.841
15	transitions easily between activities	.84		.553				.455
	Cognitive Ability (α = .85; KMO = .82)		**63.8**					
1	recognize 12 letters	.81		.651	.641			.552
2	recognize pairs of words that rhyme	.82		.789				.624
3	match letters with objects whose names start with those letters	.80		.819	.474			.675
4	name and sound the first letter in common words	.81		.870				.781
5	identify syllables by clapping hands	.82		.621	.467			.416
6	recognize numbers to 8	.82		.447	.824			.680
7	count 12 identical objects	.83			.797			.641
8	match numbers to sets of objects	.82		.426	.864			.732
	Language and Communication (α = .92; KMO = .90)		**67.6**					
1	follow two-step instructions	.91		.756				.572
2	listen to and understand stories	.90		.880				.775
3	understand instructions and questions	.90		.856				.732
4	understand action words	.92		.695				.483
5	communicate using 5-6 word sentences	.89		.889				.790
6	use pictures to tell a story	.90		.864				.746
7	convey a precise and understandable verbal message	.91		.798				.637
	Physical Development (α = .92; KMO = .80)		**51.6**					
1	copy shapes	.77		.712	.587			.586
2	copy his or her name	.81			.643			.413
3	draw a recognizable person	.79		.763				.594
4	catch a soccer ball using both hands	.80			.736			.564
5	run and kick a soccer ball	.78			.800			.641
6	jump forward several steps	.78		.770	.448			.602
7	dance rhythmically to music	.79		.762				.626
8	sufficiently energetic to participate in all class activities	.77		.685	.534			.524
9	healthy and disease-free	.80			.559			.314
10	free of physical or sensory handicaps	.80		.469	.477			.303

Concurrent and discriminant validity. Table 2 below presents the results of the Pearson correlations and their associated coefficients of determination for the correlations

between the scores on the ÉPE-AE and the ÉVIP and WIAT-II. The Awareness of Self and Environment, Cognitive Ability, and Language and Communication domains correlated most strongly with the ÉVIP and the WIAT-II, with all correlations significant at the .01 level (2-tailed). The high correlations and coefficients of determination provide strong evidence of concurrent validity for these three scales with both the ÉVIP and the WIAT-II. In contrast, although correlations between the scale scores and those of the ÉVIP and the WIAT-II are significant at the .01 level (2-tailed), only about 11% and 9% of the variance in the Social Skills and Behaviour and the Physical Development domains, respectively, can be accounted for by the ÉVIP and the WIAT-II. These small percentages are not surprising since the ÉVIP and the WIAT-II measure early reading skills rather than behaviours or physical development. The empirical data presented here provide evidence of discriminant validity for the ÉPE-AE.

Table 2. Pearson correlation coefficient (r) and coefficient of determination ($r^{2)}$ for the ÉVIP and WIAT-II results for each ÉPE-AE domain

ÉPE-AE domains	ÉVIP		WIAT-II	
	r*	**r^2**	**r***	**r^2**
Awareness of Self and Environment	.580	.340	.520	.270
Social Skills and Behaviour	.321	.103	.346	.119
Cognitive Ability	.546	.298	.649	.421
Language and Communication	.667	.445	.453	.205
Physical Development	.314	.096	.298	.089

* All correlations are significant at the .01 level (two-tailed)

Measurement bias analysis. All ÉPE-AE items were tested for language bias, the analyses for which identified by both the Mantel-Haenszel test and IRT analysis are set out in Table 3. The Mantel-Haenszel ln (estimation) is shown for each biased or DIF item and the ability range where bias is observed as obtained from the IRT DIF analysis. Finally, the favoured population is indicated. There is only one DIF item in all but the Cognitive Ability domain where three items were identified as DIF, two favouring francophones and one favouring anglophones. Items favouring anglophones and francophones across all domains were almost equal with four favouring anglophones and three favouring francophones. The ability levels over which these items showed DIF ranged from

approximately -2.5 to +0.7, which is not surprising given that the ÉPE-AE items were designed to provide the most information in this range.

Table 3. DIF items per domain

ÉPE-AE domain	DIF items	DIF analysis		
		Mantel-Haenszel ln(estimation)	Favoured group	IRT ability range
Awareness of Self and Environment	#3 identify community member's roles	+1.150	Anglophones	-2.5 – 1.0
Social Skills and Behaviour	#14 has difficulty staying attentive	+0.690	Francophones	-2.0 – 0.5
Cognitive Ability	#4 name and sound the first letter in common words	-1.096	Anglophones	-2.0 – 0.8
	#5 identify syllables by clapping hands	+3.312	Francophones	-3.0 – 1.8
	#8 match numbers to sets of objects	+2.085	Francophones	-3.0 – -0.4
Language and Communication	#4 understand action words	-1.153	Anglophones	-2.3 – 0.7
Physical Development	#8 sufficiently energetic to participate in all class activities	-1.371	Anglophones	-3.0 – 0.4

Discussion

In answering the first research question—Is the ÉPE-AE a reliable measure for assessing children's developmental status?—our investigation shows that the reliability of the complete ÉPE-AE assessment measure was excellent ($\alpha = .91$). High Cronbach coefficient alpha values were obtained in each of the five domains even though two domains, Awareness of Self and the Environment and Language and Communication, contained only seven items in their measurement scales. Results clearly indicate that the ÉPE-AE's internal consistency reliability render it a highly reliable assessment instrument. This is important since the purpose of the ÉPE-AE is to screen each francophone child at entry to kindergarten to determine whether children are at risk developmentally on five core school readiness domains. The ÉPE-AE is designed to support those who present with potential difficulties by identifying where targeted interventions and programs, and ongoing monitoring, are needed for delivery to both groups and individual students.

Study findings also provide strong validity evidence in response to research question two: To what extent does the ÉPE-AE show strong content, construct, and convergent and divergent validity? ÉPE-AE construct validity was as anticipated in four of five domains. Awareness of Self and Environment and Language and Communication, both had one principal component. Cognitive Ability had two; the first pertaining to early reading skills, and the second to mathematical skills. Physical Development also had two principal components: gross motor skills and fine motor skills. The Social Skills and Behaviour domain showed four components instead of the anticipated separate components related to each of the five scales included in this domain: hyperactivity, inattention, anxiety, emotional difficulties, and physical aggression. Instead, one combined component emerged composed of items pertaining to both physical aggression and attentiveness in class. This result was somewhat surprising and difficult to explain since one would expect these constructs to be separate. Though an explanation for this may seem unclear, Brennan, Shaw, Dishion, and Wilson (2012, p. 1290) suggest that there may be a relationship between aggression and inattention, or lack of engagement in learning. Children who act aggressively may not engage in academic learning tasks, and therefore may exhibit higher levels of inattention during learning. The second component pertained to depression and anxiety while the other two, hyperactivity and emotional difficulties, were each composed of only one item. Thus, the factor structure of Social Skills and Behaviour was not as clear as that of the other four domains. Items with double loadings were placed with the components most closely related to them conceptually.

ÉPE-AE domains showed excellent concurrent and discriminant validity with regards to the ÉVIP and the WIAT-II "gold standards." Since both tests measure early reading skills, it is not surprising that the domains addressing these skills in whole or in part—Awareness of Self and Environment, Cognitive Ability, and Language and Communication—had relatively high correlation coefficients with the "gold standards," thereby suggesting strong concurrent validity. In contrast, the Social Skills and Behaviour and Physical Development domains did not show strong correlations. These results are predictable since these domains are the least similar and thus least correlated to the language and cognition domains. Validity findings have important practical significance since they show that the ÉPE-AE assesses what it is designed to assess—five domains commonly accepted as foundational to early childhood development. Ultimately, strong construct

and concurrent and discriminant validity results indicate that the ÉPE-AE generated trustworthy data with the tested population.

In response to research question three—Does the ÉPE-AE or the EYE-TA show bias in favour of one language group over the other?—seven of the ÉPE-AE's 47 items showed DIF behaviour, four advantaging anglophone and three advantaging francophone students. Four of the five domains had only one DIF item, which greatly reduces its impact in skewing results, and in the overall interpretation of the domain's potential bias. The Cognitive Ability domain had three DIF items, one favouring anglophones and two favouring francophones. The effect of these items on possible bias in this domain is reduced since only a single DIF item remains once one of the favourable francophone items is nullified by the presence of a favourable anglophone item, both of which essentially cover the same ability range. The study's design does not provide insight into reasons for explaining the presence of DIF items. However, we can say that, due to the small number of DIF items for each linguistic population, the interpretation of the assessment's results as a whole would not be influenced in any meaningful way due to the language of the assessment. Investigating the ÉPE-AE and EYE-TA for possible assessment bias due to language is important because both versions are used not only in New Brunswick but also in several other Canadian provinces as well as internationally. Jurisdictions wishing to use both versions for comparative purposes must, and can, based on this study's findings, feel confident that the data are indeed comparable.

This study has a number of strengths and represents a significant contribution to understanding the psychometric properties of the ÉPE-AE since it is the first to quantify the instrument's internal consistency reliability, and its content, construct, and concurrent and discriminant validity in each domain it measures. Instrument bias is also assessed based on two linguistic populations and, using two independent methods, shows that the ÉPE-AE is unbiased relative to the English EYE-TA version.

However, this study also has certain limitations. Quantitative studies in general gain by having large sample sizes to increase precision and reduce sampling variability (Biau, Kerneis, & Porcher, 2008). On that basis alone, going forward with subsequent research on the ÉPE-AE, the francophone data set of 193 could be increased. Not only would the larger data set address issues of precision and sampling variability but it would also help with generating clearer subdomains in the Social Skills and Behaviour domain. Similarly, principal components analysis showed that Language and Communication had

only one component. A larger sample size combined with a confirmatory factor analysis would better enable researchers to distinguish between and test for receptive and expressive vocabulary factors. Adding items intended to contribute to these two subdomains would increase the likelihood of generating the two clearly defined subdomains. Even so, each ÉPE-AE domain generated clear principal components despite some having as few as seven items. This finding, along with excellent concurrent and discriminant validity results, suggests that the 193 sample size was sufficient for establishing important validity properties of the measure. Another encouraging result is that item response theory resulted in successful convergence of the data, thus generating item parameter estimates for discrimination, item difficulty, and pseudo-guessing for each item.

A second limitation is that the study is descriptive and not explanatory (Zumbo, 2009), which means that while we can quantify many of the assessment's psychometric properties, we know little about the effect of the assessment's context on these properties. Future research could account for these contextual properties by designing studies that compare the ÉPE-AE's psychometric properties resulting from its use in different contexts. Finally, a third limitation is the lack of data collected on teacher fidelity to the assessment's administration and whether teachers interpreted each individual item as intended. This factor is mitigated to a significant extent, however, given the extensive training conducted with the small group of 11 study teachers, during which time ongoing discussions about domain measure items along with regular checks for understanding were conducted. These processes led researchers to feel confident that teachers' understandings matched those intended in the item questions, and that study implementation could proceed. Going forward, however, future studies could include a one-to-one interview with each teacher before conducting the assessment and conduct regular checks during assessment implementation to ensure assessor understandings and the intent of the measure's items are aligned.

Future research is also needed to assess the instrument's predictive validity. It is important to know those domains that best predict future outcomes such as academic performance, the need for individualized student education plans, potential school dropout, and other outcomes measured throughout a student's academic career. Knowing the measure's predictive validity is therefore important both for early identification, and for establishing a system of ongoing assessment-based monitoring and targeted intervention to track children in need longitudinally.

To conclude, this work makes several significant contributions to early childhood monitoring and assessment research. Findings add to the literature pertaining to the early identification of vulnerable francophone students and provides curricular and assessment insights for working with children in French. This study is important because demonstrating the ÉPE-AE's strong psychometric properties, and having ruled out language bias, is crucial for francophone populations given the many challenges these students face, due largely to language development delays (Wagner, Corbeil, Doray, & Fortin, 2002). Given the importance of a strong start in school, it is important that New Brunswick, and other jurisdictions, identify vulnerable francophone students during the first months of formal schooling so learning needs can be addressed as early as possible (Canadian Language & Literacy Learning Network, 2009; CMEC, 2004; Dufour-Martel & Desrochers, 2011; University of California Davis Health System, 2009). As such, this study holds important practical implications given its purpose as an early, initial screener across multiple developmental domains that can impede a child's early and ongoing social and academic success. Since the ÉPE-AE has strong psychometric properties, it can be used with confidence to screen for at-risk children in francophone kindergartens. In turn, the relevant interventions and programs educators may implement based on ÉPE-AE results will help children who may otherwise be identified as at-risk later in life, or worse, left to suffer the consequences of not being identified at all. Ultimately, we hope that this study will influence francophone policy makers to create screening and intervention programs that ensure all kindergarten students receive the necessary help and support to which they are entitled.

References

Biau, D. J., Kerneis, S., & Porcher, R. (2008). The importance of sample size in the planning and interpretation of medical research. *Clinical Orthopaedics and Related Research*, *446*, 2282–2288. doi:10.1007/s11999-008-0346-9

Brennan, L., Shaw, D., Dishion, T., & Wilson, M. (2012). Longitudinal predictors of school-age academic achievement: Unique contributions of toddler-age aggression, oppositionality, inattention, and hyperactivity. *Journal of Abnormal Child Psychology*, *40*, 1289–1300. doi 10.1007/s10802-012-9639-2

Cabell, S., Justice, L., Konold, T., & McGinty, A. S. (2011). Profiles of emergent literacy skills among preschool children who are at risk of academic difficulties. *Early Childhood Research Quarterly*, *26*, 1–14. doi:10.1016/j.ecresq.2010.05.003

Canadian Council on Learning–Conseil Canadien de l'Apprentissage (CCL-CCA). (2006). *Lessons in learning: Why is high-quality child care essential? The link between quality child care and early learnin*g. Retrieved from EECD website: http://www.child-encyclopedia.com/sites/default/files/docs/suggestions/high-quality-child-care.pdf

Canadian Council on Learning–Conseil Canadien de l'Apprentissage (CCL-CCA). (2008). *Bringing it together: Merging community-based, life-course, linked data, and social indicator approaches to monitoring child development*. Proceedings from the Early Childhood Learning Knowledge Centre's Monitoring Committee Workshop, Toronto, ON. Retrieved from DesLibris website: http://deslibris.ca/ID/251873

Canadian Education Statistics Council. (2009). *Key factors to support literacy success in school-aged populations: A literature review*. Toronto, ON: Society for the Advancement of Excellence in Education.

Canadian Language & Literacy Learning Network. (2009). *National strategy for early literacy: Report and recommendations*. Retrieved from the Eye On Kids website: http://eyeonkids.ca/docs/files/national_strategy_for_early_literacy_report%5B1%5D.pdf

Canadian Psychology Association. (1995). *Predicting and preventing early school failure: Classroom activities for the preschool child*. Ottawa, ON: Simner, M. L.

CMEC. (2004). *Pan-Canadian results of minority Francophone students in the school achievement indicators program*. Toronto, ON: Council of Ministers of Education.

Daily, S., Burkhauser, M., & Halle, T. (2010). A review of school readiness practices in the States: Early learning guidelines and assessments. *Child Trends*, *1*(3), 1–12.

Doherty, G. (1997). Zero to six: The basis for school readiness. Technical Paper, R-97-8E, HRDC. Retrieved from the Research Gate website: https://www.researchgate.net/publication/242683154_Zero_to_Six_The_Basis_for_School_Readiness

Dufour-Martel, C., & Desrochers, A. (2011). Psychometric properties of IDAPEL (Indicateurs dynamiques d'habiletés précoces en lecture): French-language early literacy measures with students learning to read in French. Technical Report No. 12. Eugene, OR: Dynamic Measurement Group.

Dunn, L. M., Dunn, L., Leota, M., Lloyd, M. & Thériault-Whalen, C. M. (1993). *Échelle de vocabularie en images Peabody (ÉVIP)*. Retrieved from Pearson Canada website: https://www.pearsonclinical.ca/fr/products/product-master/item-75.html

Fox, L., Dunlap, G., & Cushing, L. (2002). Early intervention, positive behaviour support, and transition to school. *Journal of Emotional & Behavioural Disorders*, *10*(3), 149–158.

Garson, G. D. (2013). *Factor analysis*. Asheboro, NC: Statistical Associates Publishers.

Greenwood, C., Bradfield, T., Kaminski, R., Linas, M., Carta, J., & Nylander, D. (2011). The Response to Intervention (RTI) approach in early childhood. *Focus on Exceptional Children*, *43*(9), 1–24.

Good, R., & Kaminski, R. (2003). *Dynamic indicators of basic early literacy skills, Sixth Edition*. Longmont, CO: Sopris West Educational Services.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology*, *123*, 207–215.

Huck, S. W. (2008). *Reading statistics and research* (4th ed.). Boston, MA: Pearson Education.

Invernizzi, M., Sullivan, A., Meier, J., & Swank, L. (2004). *Phonological awareness literacy screening: Preschool*. Charlottesville, VA: University of Virginia.

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), 1–22. doi:10.1037/cjbs2007001

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151. doi:10.1177/001316446002000116

KSI Research International. (2009). *Validity and reliability of the EYE-TA*. Retrieved from EYE website: https://www.earlyyearsevaluation.com/images/site_docs/Reliability_Validity_EYE-TA.pdf

Lyon, G., Fletcher, J., Shaywitz, S., Shaywitz, B., Torgeson, J., Wood, F., Schulte, A., & Olson, R. (2001). Rethinking learning disabilities. In C. Finn, A. Rotherham, & C. Hokanson (Eds.), *Rethinking special education for a new century* (pp. 259–287). Washington, DC: The Thomas B. Fordham Foundation.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

McCartney, K. (2007). Current research on child care effects. Retrieved from EECD website: http://www.child-encyclopedia.com/sites/default/files/textes-experts/en/857/current-research-on-child-care-effects.pdf

National Education Goals Panel. (1991). The Goal 1 Technical Planning Subgroup report on school readiness. In *Potential strategies for long-term indicator development. Reports of the technical planning subgroups*. Washington, DC: National Education Goals Panel. Retrieved from ERIC website: http://files.eric.ed.gov/fulltext/ED350340.pdf

National Governors Task Force on School Readiness. (2005). *Building the foundation for bright futures: Final report on the NGA task force on school readiness*. Washington, DC: National Governors Association. Retrieved from NGA website: http://www.nga.org/files/live/sites/NGA/files/pdf/0501TASKFORCEREADINESS.pdf

National Research Council (NRC). (2008). Early childhood assessment: Why, what, and How. Committee on Developmental Outcomes and Assessments for Young Children, C.E. Snow and S.B. Van Hemel, Editors. Board on Children, Youth, and Families, Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National School Readiness Indicators Initiative. (2005). *Getting ready: Findings from the National School Readiness Indicators Initiative: A 17 state partnership*. Providence, RI: Rhode Island, KIDS COUNT. Retrieved from Getting Ready website: http://www.gettingready.org/matriarch/d.asp?PageID=303&PageName2=pdfhold&p=&PageName=Getting+Ready+-+Full+Report%2Epdf

Nunnaly, J. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Oshima, T. C., & Morris, S. B. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, *27*(3), 43–50. doi:10.1111/j.1745-3992.2008.00127.x

Sloat, E. A., Beswick, J. F., & Willms, J. D. (2007). Using early literacy monitoring to prevent reading failure. *Phi Delta Kappan*, *3*, 523–529.

Stedron, J., & Berger, A. (2010). *NCSL technical report: State approaches to school readiness assessment*. National Conference of State Legislators. Retrieved from NCSL website: http://www.ncsl.org/documents/Educ/KindergartenAssessment.pdf

Tabachnik, B., & Fidell, L. (2001). *Using multivariate statistics* (4th ed.). Toronto, ON: Allyn & Bacon.

The Learning Bar. (2016). *Early years evaluation*. Retrieved from The Learning Bar website: http://thelearningbar.com/solutions/home-school-transition/early-years-classrooms/

Thordardottir, E., Keheyia, E., Lessard, N., Sutton, A., & Trudeau, N. (2010). Typical performance on tests of language knowledge and language processing of French-speaking 5-year-olds. *Canadian Journal of Speech-Language Pathology and Audiology*, *34*(1), 5–16.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, 469–477. doi:10.1037/h0071663

University of California Davis Health System. (2009). Poor attention in kindergarten predicts lower high school test scores. *Science Daily*. Retrieved from UCDMC website: http://www.ucdmc.ucdavis.edu/medicalcenter/features/2009-2010/09/20090924_attention_span.html

Vasilyeva, M., Ludlow, L. H., Casey, B. M., & St. Onge, C. (2009). Examination of the psychometric properties of the measurement skills assessment. *Educational and Psychological Measurement*, *69*(1), 106–130. doi:10.1177/0013164408318774

Wagner, S., Corbeil, J.-P., Doray, P., & Fortin, É. (2002). Alphabétisme et alphabétisation des francophones au Canada. Ottawa, ON: Statistiques Canada et Département des ressources humaines Canada.

Wechsler, D. (2001). *Weschler Individual Achievement Test – Second Edition (WIAT-II)*. Retrieved from Pearson website: http://www.pearsonclinical.com/psychology/products/100000664/wechsler-individual-achievement-test-second-edition-wiat-ii.html

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

Zwyno, M. S. (2003). *A contribution to validation of score meaning for Felder-Soloman's index of learning styles*. Proceedings of the 2003 American Society for Engineering Education Annual Conference & Exposition. American Society for Engineering Education. Retrieved from: http://www4.ncsu.edu/unity/lockers/users/f/felder/public/ILSdir/Zywno_Validation_Study.pdf