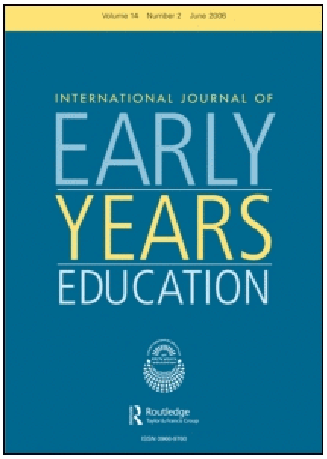

International Journal of Early Years Education
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713425018

Are the indicators for the Language and Reasoning Subscale of the Early Childhood Environment Rating Scale-Revised psychometrically appropriate for Caribbean classrooms?

Michael Canute Lambert [a]; Sian G. Williams [b]; Johnetta W. Morrison [a]; Maureen E. Samms-Vaughan [c]; Wayne A. Mayfield [a]; Kathy R. Thornburg [a]
[a] University of Missouri-Columbia, Missouri, USA
[b] Early Childhood Consultant, Jamaica
[c] University of the West Indies-Mona, Jamaica

Online Publication Date: 01 March 2008

To cite this Article: Lambert, Michael Canute, Williams, Sian G., Morrison, Johnetta W., Samms-Vaughan, Maureen E., Mayfield, Wayne A. and Thornburg, Kathy R. (2008) 'Are the indicators for the Language and Reasoning Subscale of the Early Childhood Environment Rating Scale-Revised psychometrically appropriate for Caribbean classrooms?', International Journal of Early Years Education, 16:1, 41 — 60

To link to this article: DOI: 10.1080/09669760801892219
URL: http://dx.doi.org/10.1080/09669760801892219

International Journal of Early Years Education
Vol. 16, No. 1, March 2008, 41–60

Are the indicators for the Language and Reasoning Subscale of the Early Childhood Environment Rating Scale-Revised psychometrically appropriate for Caribbean classrooms?

Michael Canute Lambert[a]*, Sian G. Williams[b], Johnetta W. Morrison[a], Maureen E. Samms-Vaughan[c], Wayne A. Mayfield[a] and Kathy R. Thornburg[a]

[a]*University of Missouri-Columbia, Missouri, USA;* [b]*Early Childhood Consultant, Jamaica;* [c]*University of the West Indies-Mona, Jamaica*

Evaluating the psychometric properties of the indicators that comprise the Early Childhood Environment Rating Scale-Revised (ECERS-R) language-reasoning scale from an item response theory (IRT) perspective on a sample of observations from 334 Caribbean classrooms, Stout's procedure revealed that all indicators on this dimension are not part of a single essentially unidimensional construct. IRT-based factor analyses on the indicator scores yielded two factors – named *Language-Reasoning Activities* and *Language-Reasoning Materials*. IRT analyses conducted on these two factors revealed that their indicators provide adequate psychometric information and have no floor effects – although they demonstrate evidence for ceiling effects. IRT also revealed that at least within the Caribbean context: (a) the ECERS-R authors have ordered the indicators inappropriately; (b) administration of all indicators is unnecessary; and (c) equally weighting indicators might yield spurious results. IRT-based scoring might improve the psychometric soundness of indicators on this ECERS-R scale.

Keywords: Caribbean; early childhood classrooms; ECERS-R; item response theory

ECERS-R

The Early Childhood Environment Rating Scale (ECERS; Harms and Clifford 1980) and the ECERS-R (Harms, Clifford, and Cryer 1998), its revised edition, are measures used to assess the quality of environments within early childhood settings such as preschool classrooms (Harms, Clifford, and Cryer 1998; Sakai et al. 2003). The ECERS-R is widely used in North America and other parts of the world to assess programme quality in classrooms serving children ages 2.5 through 5 years (Tietze et al. 1996; Harms, Clifford, and Cryer 1998; Clifford et al. 2005; Goelman et al. 2006). The purposes of its use include programme monitoring, improvement, and evaluation as well as early childhood environment quality-based research (Farquhar 1989; Calder 1996; Tietze et al. 1996; Harms, Clifford, and Cryer 1998; Aboud 2006).

ECERS-R in International Context

To make its use more appropriate within an international context, some of the ECERS-R items have been changed to make them more culturally appropriate for specific groups (see Beller et al. 1996; Lee, Lee, and Lee 1997). In addition, the authors of the measures (Harms, Clifford, and Cryer 1998) have reported that the ECERS-R has been translated into several languages, including German, Swedish, Icelandic, Portuguese, Italian and Spanish. The instruments have been used in countries within Asia and Europe (Herrera et al. 2005; Aboud 2006; Goelman et al. 2006).

*Corresponding author. Email: lambertmc@missouri.edu

ISSN 0966-9760 print/ISSN 1469-8463 online
© 2008 Taylor & Francis
DOI: 10.1080/09669760801892219
http://www.informaworld.com

Although not directly related to the use of ECERS-R in international contexts, recent research has shown that early childhood professionals in different international settings tend to agree on what contributes toward positive early childhood development (Tietze et al. 1996). Considered with research findings on the ECERS-R in North America documenting an association with child developmental outcomes (see Burchinal et al. 2002; Moore et al. 2002), the cross-national convergence of views might indirectly suggest promise for the criterion-related validity in the use of the ECERS-R in international context, especially as it pertains to child development outcomes.

Despite efforts to make its international use more appropriate, use of the ECERS-R in such contexts has been criticized. One criticism is that simply translating this measure into different languages is insufficient to justify its use within other countries. This criticism might be justified since most psychometricians and the professional organisations (e.g. the International Commission) to which they belong have cautioned that it is inappropriate to use a measure (e.g. the ECERS-R) that was designed specifically for the assessment of constructs in one nation, to assess such constructs in other nations. More specifically, as far as early childhood education is concerned, customs, practices and needs might differ according to educational philosophy, definitions of quality, and the contexts in which early childhood education occurs (Calder 1996). Thus, it has been argued that it is inappropriate to use the ECERS-R to examine early childhood environment quality in societies whose cultural mores and concomitant child-rearing and childcare philosophies, and practices, might differ from those of the United States, the country in which the ECERS-R was developed (Karrby and Giota 1994; Tietze et al. 1996).

The concerns detailed in the preceding paragraphs are also evident in the use of the ECERS-R in the Caribbean, where more recently it was used to assess classroom quality in multiple English-speaking Caribbean nations. This use occurs in spite of the absence of information on the psychometric soundness of its use in international contexts. Absence of information on the psychometric properties for the ECERS-R in the Caribbean can be consequential for environment quality assessment and findings derived from research that focuses on programme quality in early childhood settings within Caribbean countries. It is, therefore, difficult to interpret findings from Caribbean classrooms since they could be reflective of measurement artifacts and not of the true quality levels evident in Caribbean early childhood settings.

Besides concerns regarding the use of the ECERS-R in international contexts, it is important to note that this measure was developed using methodology that was guided by classical test theory (CTT). CTT-based studies conducted in the United States have documented that the ECERS-R possesses adequate psychometric properties (Vandell and Wolfe 2000). For example, large numbers of research projects have provided evidence of appropriate content and criterion-related validity, acceptable interrater reliability, and high internal consistency for its subscales (Scarr, Eisenberg, and Deater-Decker 1994; Clifford et al. 2005; Goelman et al. 2006). Despite evidence that supports the psychometric soundness of the ECERS-R, we have found no studies that have empirically verified the subscale structure of the instrument. Therefore, no factor analytic studies of the ECERS-R have replicated the seven-factor model the original subscales purported. Several studies reported finding only one overall factor (Scarr, Eisenberg, and Deater-Decker 1994; Helburn 1995; Phillipsen et al. 1997). Other studies found evidence for two underlying factors (Peisner-Feinberg et al. 2001; Burchinal et al. 2002; Cassidy et al. 2005; Clifford et al. 2005) that are relatively similar across studies, including a methodologically rigorous study, where a two-factor model derived from exploratory factor analysis on a derivation sample was later confirmed on a cross-validation sample using confirmatory factor analysis (Cassidy et al. 2005).

Since CTT-driven methodology dictates its psychometric properties, ECERS-R indicators[1] are administered to determine each item score and all items are routinely scored when the ECERS-R is used in evaluating preschool settings (see the next paragraph and 'Description of measures' in the Method section for the distinction between indicators and items). In addition, although the scale has 43 items, each item is scored on the basis of ratings derived from 10 or more indicators, resulting in a total of 470 indicators. A single ECERS-R scale might therefore have hundreds of indicators from which its score is derived. The ECERS-R authors have ordered the indicators for each item based on face validity, but, to our knowledge, these indicators have not been pre-calibrated. Thus, they have not been submitted to quantitative analyses that verify that their order is valid.

Since scores for each of the 43 ECERS-R items are derived from scores on groups of indicators that are not calibrated, the indicators might have implications for precision of the ECERS-R items in measuring the quality levels they purportedly measure. One concern is that we cannot be certain that the indicators are ordered appropriately. Inappropriately ordered indicators might have major effects on item accuracy. Moreover, each item is scored by totaling the number of positively scored indicators that are ordered under specific points in their respective item's Likert scale. Thus, when administered, each indicator under each Likert scale point is scored equally. Yet no information on the appropriateness of this practice exists. An even more fundamentally important implication is that there is no empirical evidence that the dichotomous indicators for each scale satisfy key measurement assumptions including unidimensionality. Thus, considered within the context of the nation in which the ECERS-R was developed and in the Caribbean context in which it is now used, further investigation is needed to address the soundness of such use.

Beginning to address the concerns detailed above, the present study focuses on data from early childhood settings aggregated across two English-speaking Caribbean nations, Grenada and Jamaica. Focusing on the dichotomous indicators of the language-reasoning scale, one of the seven ECERS-R scales that has been documented as being associated with developmental outcomes in preschoolers (Burchinal et al. 2000; NICHD Early Child Care Research Network 2000), this study addresses the following six objectives: (a) to determine whether dichotomous indicators of the language-reasoning scale are part of an essentially unidimensional scale and to take steps to determine the most appropriate factor model evident in the data; (b) to evaluate whether the indicators from items on the scale(s) (derived from objective [a]) provide sufficient information to warrant retention on their respective scales or whether some items might require further study – that is, to determine whether the indicators appropriately discriminate in measuring the construct of interest; (c) to learn whether it is appropriate to administer all indicators from the factor on which they load, regardless of the quality levels of the early childhood setting being assessed; (d) to determine whether applying equal weighting to each indicator on each factor is appropriate; (e) to discover whether indicators appropriately measure early childhood environment settings with varying quality levels – in other words, to learn whether indicators on each scale are capable of assessing the quality of early childhood settings that might have poor, adequate, and superior quality on the domains assessed; and (f) to learn whether the manner in which the ECERS-R authors have ordered indicators as measuring low to high quality is appropriate.

Method

Sample

A total of 334 early childhood classrooms from the nations of Jamaica and Grenada were sampled. Of this number 200 were Jamaican classrooms and the remaining classrooms were Grenadian. The observations were a part of research projects conducted from 2005 to 2006 to assess the quality of early childhood settings within the Caribbean Basin. Other nations were sampled but data on the ECERS-R indicators were unavailable at the time this study was conducted.

Data collection procedures

Country teams of between 5 and 9 experienced early childhood practitioners engaged in a three-day training programme focused on appropriate administration of the ECERS-R. In each country the assessment procedure was piloted in two early childhood centres visited by trained observers in groups of two or three. After each visit, the ratings were compared and differences between raters discussed in the team, led by a senior researcher with extensive training and experience in the administration of the ECERS-R. This process identified common understanding of all indicators but especially those that needed clarification in local contexts (e.g. the convention of children resting by laying their heads on their desks, which was agreed as inadequate; the provision of food and snacks from home, which was treated in the same way as those provided by the centre from the point of view of nutritional content received by the child). Each centre was observed and rated over the course of a preschool day, which in most Caribbean countries is an extended morning (8.00 a.m. to 2.30 p.m. is typical). It is important to note that because all indicators on the ECERS-R are scored, the observation time in the Caribbean nations is considerably lengthier than the four-hour observation window often employed in US classrooms, where only indicators needed to achieve an item score are administered (see 'Scoring procedures' below). There were very few differences between raters after the visit to the second centre in the pilot study.

For the administration of the ECERS-R, observers visited centres either in pairs or on their own, meeting together once weekly with a survey team leader. The senior researcher provided technical guidance at the survey team meetings by telephone and ensured consistency in application of the instrument. The score sheets were also checked individually by the senior researcher to ensure accuracy. Inconsistencies (generally associated with indicators written as negative statements) were addressed immediately and observers asked to re-visit the centre they previously assessed to ensure their ratings were accurate. In instances where an observer made routine errors after the first team meeting, the observer's centres were re-visited and rated by another observer. In each country, one member of the team was unable to use the instrument consistently and the centres they observed were re-visited for confirmation. The senior researcher checked all items on each sheet and did the final scoring.

Measure

General description

The ECERS-R is an observation-based measure of early childhood environment quality with 43 items grouped under seven subscales. As described below, scores for each of these 43 items are derived from rating patterns on multiple dichotomously scored indicators.[2] The seven subscales are space and furnishings, personal care routines, language-reasoning, activities, interaction, programme structure, and parents and staff.

Scoring procedures

Each item on the ECERS-R has multiple dichotomously scored indicators that are scored *Yes/No* and, in some instances, *Yes, No, Not Applicable (N/A)*. These indicators represent four points along the 1 to 7 scoring continuum – 1, 3, 5, and 7. The first set of indicators at 1 presents descriptions and examples for an item's content that are *inadequate*. The second set of indicators is associated with a score of 3 (*minimal*), the third set with a score of 5 (*good*), with the final indicators associated with a top score of 7 (*excellent*). Based on these indicators, observers assign a score from 1 to 7 for each item. A rating of 1 is given if any of the indicators under 1 is scored *Yes*.

A rating of 2 is given if all indicators under 1 are marked *No* and at least half the items under 3 are scored *Yes*. A rating of 3 occurs if all indicators under 1 are scored *No* and all indicators under 3 are scored *Yes*. A rating of 4 is awarded if all indicators under 3 are scored *Yes* and at least half of the indicators under 5 are scored *Yes*. A rating of 5 is given when all indicators under 5 are met. A rating of 6 is given when all indicators under 5 are met, as well as half the indicators under 7. A rating of 7 occurs when all indicators under 7 are scored *Yes*. In some instances, *N/A* may be scored for some indicators and items (mostly indicators and items that address serving children with disabilities). Mean subscale scores are calculated by summing individual item scores and dividing by the total number of items scored; the total ECERS-R score equals the sum of all individual items divided by the total number of items scored. We note that in most research and evaluation settings, all the dichotomous indicators are rarely administered since once the criterion or criteria are met for scoring a specific item, raters move on to the next items without scoring the rest of the indicators.

Reliability

The ECERS-R authors reported interrater reliability/agreement of 86% across all 470 indicators. The authors also report that, at the item level, the proportion of agreement was 48% for exact agreement and 71% for agreement within one point. With respect to internal consistency, the ECERS-R authors reported coefficient alphas ranging from .71 to .88 for the seven original subscales and .92 for the total scale. Perlman, Zellman, and Lee's (2004) examination of correlations among the individual items on the ECERS-R included an average inter-item correlation of .39, and item-total correlations ranging from .35 to .76.

Validity

Although there are some inconsistencies in research findings, there is evidence supporting criterion-related validity for the ECERS and ECERS-R. Holloway et al. (2001) stated that construct validity is evident for ECERS-R. Predictive validity is also evident since ECERS-R scores are often significantly correlated with child outcomes (see Burchinal et al. 2002; Moore et al. 2002). Several studies have shown that scores on the ECERS and ECERS-R are associated with structural (e.g. teacher training and education) and global classroom quality (see Cassidy et al. 1995; Phillipsen et al. 1997; Phillips et al. 2000; Burchinal et al. 2002; Perlman, Zellman, and Lee 2004; Warash, Markstrom, and Lucci 2005). Convergent validity is also evident since Sakai et al. (2003) reported that ECERS-R scores are highly correlated with scores from the *Caregiver Interaction Scale* (Arnett 1989). McCarty, Abbott-Shim, and Lambert (2001) also identified studies that found significant correlations between the ECERS and the *Assessment Profile for Early Childhood Programs: Research Version.*

Data analyses

Overview of data analyses and IRT models

Focusing on the indicators of the language-reasoning scale, item response theory (IRT) procedures were used to address the objectives of the study. Thus, the data analyses focused on IRT-based factor analyses to determine the number of factors that are represented in the indicator score, testing IRT assumptions and the estimation of the parameters for the dichotomous indicators. To make the discussion of data analytical procedures more relevant to the measure on which this study focuses, we discuss its principles and procedures within the context of early childhood environment quality.

Description of IRT and IRT models

Despite variations in types of models, most IRT models infer one or more latent variables (i.e. traits, factors, constructs) measured by observed responses (i.e. items, indicators). All IRT models describe the probability of particular rating/response to various indicators measuring a specific trait level (i.e. quality of the early childhood environment for the ECERS-R) labeled θ. In the case of ratings on indicators from the ECERS-R, IRT refers to the probability that an early childhood setting with a specific quality level would receive positive ratings for indicators that measure this level than to indicators measuring other levels of quality (Panter and Reeve 2002).

Widely used IRT models include one- (i.e. Rasch or 1PL), two- (2PL), and three-parameter (3PL) models. In the 1PL model, items are assumed to discriminate equally across various quality levels and the focus is on the level of quality of the early childhood classroom being measured. Therefore, only the location parameter estimate (*b)* reflecting this quality level for an indicator on a specific ECERS-R scale is estimated. The two-parameter model does not assume equal discrimination for all items. Besides the *b* parameter, a discrimination parameter (*a*) is also estimated. The three-parameter model is most often used in educational settings, where a third (guessing) parameter (*c*) is estimated. The *c* parameter represents the probability of a rater giving positive ratings on indicators that measure higher quality levels when assessing a lower quality etting.

Many researchers believe that IRT item parameter estimates are far more informative than estimates obtained from CTT-based methodology. In a 2PL model, the *a* and *b* parameter estimates, for example, are often used to plot logistic trace lines known as operating characteristic curves (OCCs). IRT practitioners often view graphic representation as being an especially important benefit of IRT since visually inspecting them can lead to interpretation of the discrimination qualities each indicator affords and the quality level at which an indicator best discriminates (Marshall et al. 2002). Higher *a* parameter estimates result in steeper OCCs, whereas higher *b* parameter estimates result in shifting the location of the OCCs further to the right, indicating that only early childhood settings with relatively high quality levels are likely to receive a positive rating for this item. The *a* and *b* parameter estimates can also be used to plot an item information curve (IIC), where the *a* parameter estimate determines its height (i.e. the amount of information the item provides) and the *b* parameter estimate determines its location (i.e. the quality levels where the indicator provides the most information). The IIC is useful because it can highlight items that provide limited measurement information and can thus be targeted for further study or elimination (Hambleton, Swaminathan, and Rogers 1991).

IRT also has the ability to determine whether each indicator within a given dimension measures the same quality levels for a particular group of early childhood classrooms. This quality of the IRT methodology is especially important since by simply totaling indicator scores according to the item and subsequently the rating scale on which such items load, a common CTT practice, professionals often ignore the differences in quality levels an indicator provides. Thus, in most IRT models the ECERS-R indicators (or the items that are derived from them) would almost never be totaled since scores would be derived from examining the pattern of ratings that are given for the early childhood setting being assessed. Early childhood settings with the same observed score may have different standings on the quality level being measured because the pattern of ratings for one centre with a specific score might reflect high quality levels, while settings with the same total score, but a different pattern of ratings, might reflect low quality levels.

IRT assumptions

To obtain trustworthy item parameter estimates, four assumptions must be met. The first assumption is that of appropriate dimensionality. This assumption underscores that the procedure used to identify dimensionality (e.g. factor analyses) is conducted in a manner that minimizes spurious factor solutions and that indicators comprising each factor identified are measuring a single theoretical construct or at least measuring a dominant dimension (Stout 1990). The second is conditional independence (also known as local independence; see Thompson and Pommerich 1996), meaning that if the quality level measured is held constant there should be no associations among indicators. The third assumption is that the construct of focus (in this case the quality level of early childhood settings) is normally distributed in the population, a principle that is incorporated in the algorithm that estimates IRT parameters. Finally, the fourth assumption is that the chosen model fits the data better than other models with fewer or more parameter estimates.

Data analytical procedures for testing testable IRT assumptions

Appropriate dimensionality

Establishing the unidimensionality of constructs is crucial in IRT analyses. Although confirmatory factor analysis (CFA) is often used to demonstrate that a group of items measures a unidimensional construct, the strictness of the assumptions applied in these procedures has been criticised since most researchers acknowledge that the assumption of complete unidimensionality of factors used in the social sciences is often unrealistic. Several researchers (e.g. Stout 1990, 2005; Nandakumar 1993; Hattie 1996; Stout et al. 1996) have theorised and demonstrated that the presence of one dominant dimension in a group of items purported to measure a single dimension is not only sufficient but more practical than the strict assumption of unidimensionality. Stout has used the term *essential unidimensionality* to describe this assumption. He developed the DIMTEST software application to test for essential dimensionality (see Stout 2005) in measures whose items are rated on dichotomous scales. His procedure provides a T-statistic. A nonsignificant T indicates the presence of essential unidimensionality, whereas a significant T indicates the absence of a single dominant factor. It is important to note that this procedure requires approximately 20 (and preferably more) items/indicators in a dimension and relatively large samples for accurate estimates (see Stout 2005).

Analyses to determine factor structure of the language-reasoning scale

Multiple approaches are available to identify factors from indicators in a dataset and to assess appropriate dimensionality. Many assess this assumption by examining the intercorrelations among items. These procedures might be inappropriate when item scores result in a skewed distribution (e.g. for dichotomous items where some items are frequently or infrequently endorsed) and could result in anomalous factor models. This study used TESTFACT (Wilson et al. 2003), a software application that can conduct full information factor analysis (FIFA) from dichotomously scored items by applying multidimensional IRT models.[3] Instead of using item intercorrelations to extract factors, FIFA examines response patterns on dichotomously scored items to identify latent variables and is more likely to accurately identify multiple factors when they do indeed exist than some other existing methods (Panter and Reeve 2002). Besides FIFA item loadings, TESTFACT also provides well-recognised indices of fit (e.g. size of eigenvalues).[4]

Conditional independence

TESTFACT also has the capability of using a bi-factor data analytical procedure to test whether conditional independence is evident in a specific scale (Gibbons and Hedeker 1992).[5]

Appropriate IRT model

The purpose of the second set of analyses was to determine whether a one-, two-, or three-parameter model would provide the best fit for indicator ratings. This assumption was tested using MULTILOG (Thissen, Chen, and Bock 2003). This application is capable of testing multiple IRT models. MULTILOG calibrates item parameter estimates using marginal maximum likelihood (MML) estimation and uses a G^2 statistic to assess model fit. To test which model provided the best fit, nested models were examined where the G^2 statistics (distributed as χ^2) of 1PL, 2PL, and 3PL models were compared.

Results

Testing IRT assumptions

Testing for essential unidimensionality of all indicators

Addressing objective (a) by determining whether the indicators on the language-reasoning scale are indicators of an essentially unidimensional global factor, we submitted the indicators to a test on essential unidimensionality using the DIMTEST application. Stout's *T* statistic was significant at $p < .02$, revealing that the hypothesis of one dominant dimension for all indicators was not supported. We note that all negatively worded indicators (i.e. listed under the first point of the Likert scale of their respective items) were excluded from these and other analyses because they almost always had ratings of 1 (i.e. they were positively rated for ≤ 1% of all centres surveyed). Inclusion of these items and some one-dozen other indicators (i.e. those that almost always received ratings of 1 and were therefore excluded) risked biased parameter estimates.

Analyses to determine factor structure of the language-reasoning indicators

Continuing to address objective (a), to determine the factor structure for the indicators of the language-reasoning scale, TESTFACT nonadaptive full information factor analyses (FIFA) procedures with varimax and promax rotations (see Gibbons and Hedeker 1992) were conducted on ratings given on the indicators.[6] A theoretically plausible two-factor model emerged: the first factor was comprised of indicators that reference activities that facilitate language and reasoning and the second factor had indicators that reference materials used to promote language and reasoning. Because the indicators loading for the first and second factors in this model reflected activities and materials that promote language and reasoning respectively, the factors are labeled Language and Reasoning Activities and Language and Reasoning Materials. The loadings forming the FIFA two-factor model are listed in Table 1. Indicators that had loadings that were ≥ .3 were deemed as loading on their respective factors and were included in IRT analyses. It is important to note that two items cross-loaded on the two factors and were excluded from further analyses.

Conditional independence

TESTFACT analyses revealed that conditional dependence for the two-factor model is unjustified (Gibbons and Hedeker 1992).[7]

Appropriateness of chosen IRT model

For each factor considered separately, the MULTILOG analyses used to test whether the 1PL, 2PL, or 3PL model best represented the data revealed that the 2PL model was most appropriate for further analysis.[8]

Table 1. Factor loadings from full information factor analysis of indicators from language and reasoning scale of the Early Childhood Environment Rating Scale Revised Edition.

Brief indicator description	Language and Reasoning Activities	Language and Reasoning Materials
15.3.1. Some books	.07	**.44**
15.3.2. Receptive language	.24	**.44**
15.5.1. Book selection	.09	**.36**
15.5.2. Additional language	.23	**.42**
15.5.3. Books organised	.07	**.48**
15.5.4. Books appropriate	.15	**.48**
15.5.5. Staff read*	**.30**	**.40**
15.7.1. Books rotated	.16	**.44**
15.7.2. Books relate	.21	**.41**
16.3.1. Some activities	**.32**	.29
16.3.2. Materials accessible	.21	**.34**
16.5.1. Communication activities	**.39**	.28
16.5.2. Materials encourage communication	.28	**.41**
16.7.1. Encourage reasoning	**.40**	.12
16.7.2. Concepts respond to interest	**.40**	.27
17.3.1. Staff sometimes talk logical	**.31**	.26
17.3.2. Some concepts age appropriate*	**.37**	**.32**
17.5.1. Staff talk logical	**.30**	.24
17.5.2. Children encouraged to talk	**.30**	.22
17.7.2. Concepts match children's needs	.26	.17
18.3.2. Children allowed to talk	**.48**	.11
18.5.1. Staff conversations	**.39**	.15
18.5.2. Language used to exchange info	**.43**	.17
18.5.3. Staff add info to expand	**.33**	.18
18.5.4. Staff encourage communication	**.50**	.03
18.7.1. Staff have conversation	**.44**	.09

Note: Indicators numbers are presented as they appear in the ECERS-R manual; Factor loadings are **boldfaced**; *Not included in further analyses since they are cross-loaded.

Objective (b): evaluating measurement precision in indicators

In addressing objective (b), whether all indicators provide enough precision to be used in the assessment of the two language and reasoning factors, we examined the IIC for each indicator. Figures 1 and 2 include examples of indicators with varying information levels.[9] We focus on the first factor as a detailed example. It is important to note that the *a* parameter estimates were relatively high for all indicators, demonstrating that the indicators provide high precision levels for the quality levels they measure. Likewise, the examples of the curves in the figures reveal relatively high information curves. Since the patterns for all item information curves matched those for all indicators presented in the figures we judged that all indicators from each factor provide appropriate levels of information.

Objective (c): appropriateness of administering all indicators

Objective (c) focused on whether it is appropriate to administer all indicators in each of the two newly developed language-reasoning subscales. To address this objective, it is essential to recall

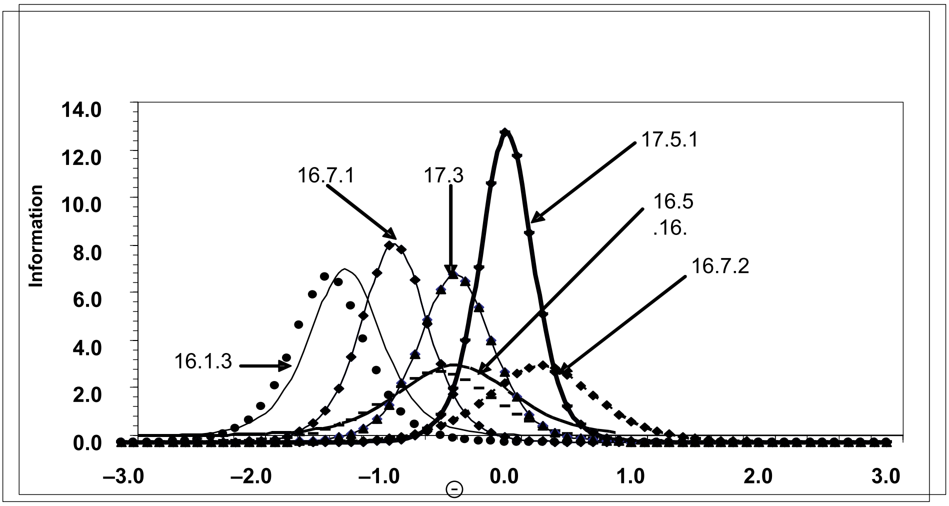

Figure 1. Examples of information curves for items with varying information levels and depicting items measuring varying quality levels for Factor 1.

that the *b* parameter estimates represent the spacing of indicator ratings on the quality continuum. That is, they represent the quality levels each indicator measures. It is important to note that like standardised scores (e.g. *z*-scores), the *b* parameter estimates are in standardised units with a mean of 0. Looking at Figures 1–4, which show indicators from the Language and Reasoning Activities and Language and Reasoning Materials scales, we first note that different indicators depicted in Figures 1 and 2 show that the information curves peak at different points on the scale. For example, Figure 1 shows that indicator 17.5.1 measures far higher quality levels than 16.1.3 for the Language Activities scale. Similar patterns were also evident in the item characteristic curves in Figure 3, where indicators differed in depicting the quality levels they measure. The *b* parameter

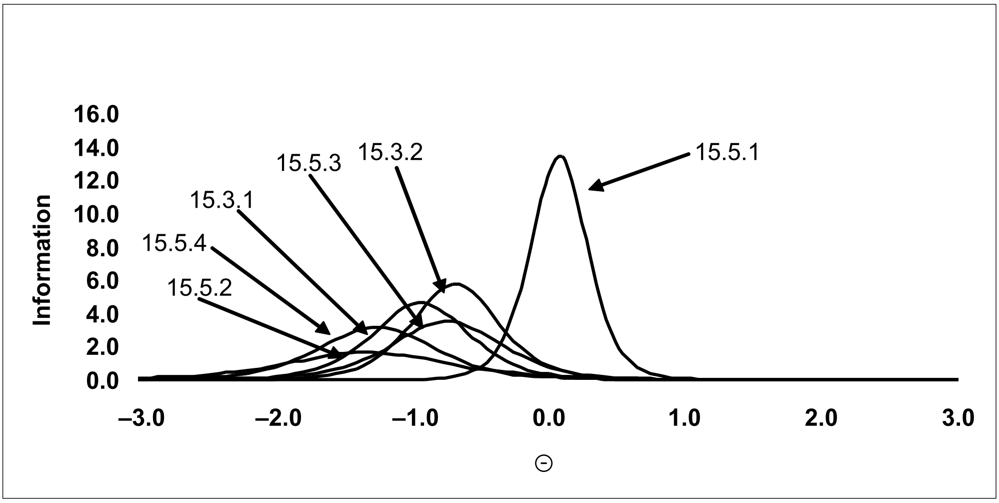

Figure 2. Examples of information curves for items with varying information levels and depicting items measuring varying quality levels for Factor 2.

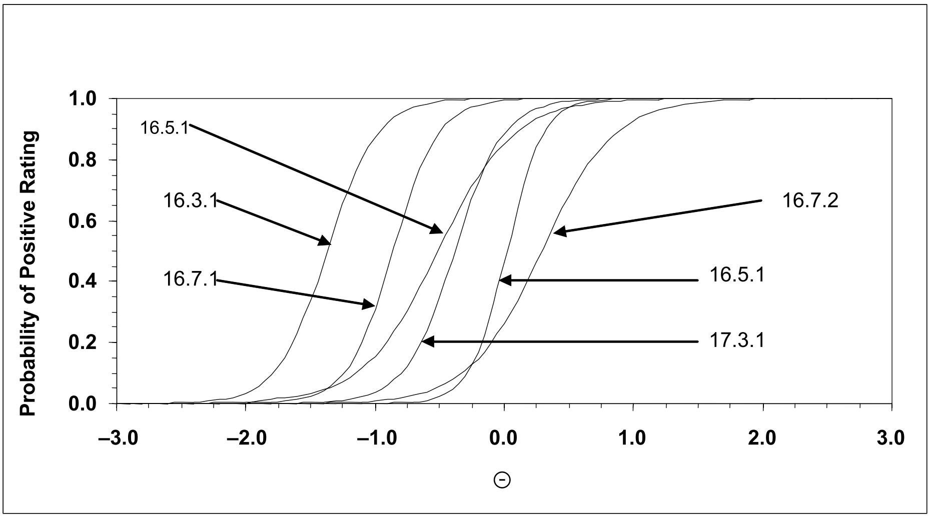

Figure 3. Item response function for item examples on Factor 1.

estimates in Table 2 also showed that 17.5.1 measures quality levels slightly above the mean, whereas 16.3.1 measures levels that are more than 1 standard deviation (SD) units below the mean. Similar patterns were evident for 15.5.1 and 15.3.1 for Language and Reasoning Materials. Thus, for classrooms that score at average levels on the Language and Reasoning Activities scale, indicator 16.3.1 does not provide as reliable a measure of quality as indicator 17.5.1 does.

Objective (d): appropriateness of totaling indicators to derive item score

Since IRT analyses demonstrate that ECERS-R indicators differ in their ability to accurately measure various quality levels, we can now address objective (d) – whether it is appropriate to

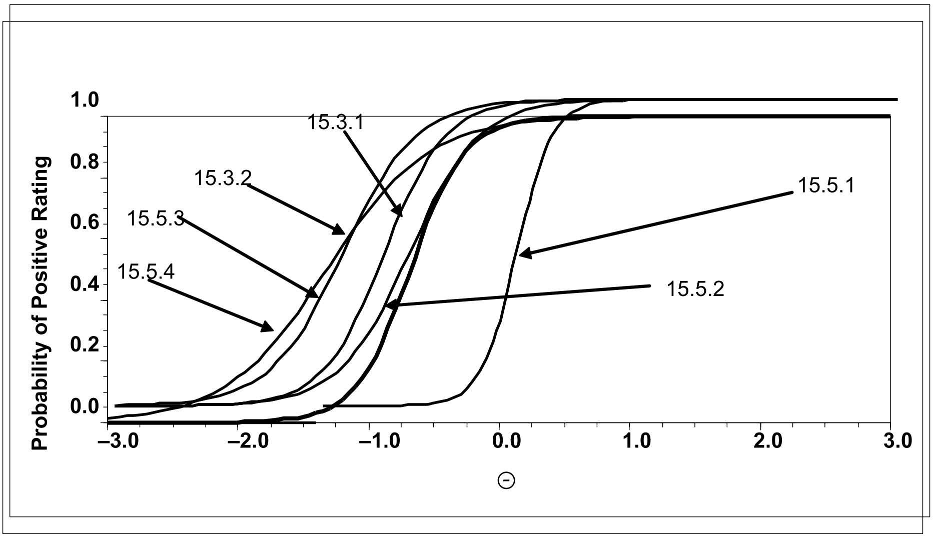

Figure 4. Item response function for item examples on Factor 2.

Table 2. Parameter estimates for language and reasoning activities.

Brief indicator description	*a*	*b*
16.3.1. Some activities	3.10	–1.37
16.5.1. Communication activities	2.02	–0.51
16.7.1. Encourage reasoning	3.40	–0.86
16.7.2. Concepts respond to interest	2.11	0.29
17.3.1. Staff talk logical relationships	3.13	–0.38
17.5.1. Staff talk logical sequencing	4.25	0.02
17.5.2. Children encouraged to talk	5.18	0.16
18.3.2. Children allowed to talk	1.42	–1.77
18.5.1. Staff conversations	2.88	–0.32
18.5.2. Language used to exchange info	2.73	–0.88
18.5.3. Staff add info to expand	3.61	–0.29
18.5.4. Staff encourage communication	1.44	–0.22
18.7.1. Staff have conversation	2.28	–0.18

Note: *a* = discrimination parameter estimates; *b* = discrimination parameter estimates.

total the scores for each indicator to provide an item score and subsequently a scale score. Indicators on a dimension with identical middle digits (e.g. 16.3.1, 17.3.1 and 18.3.2) receive identical scores in the current ECERS-R scoring procedure. If it is appropriate to merely total scores on each indicator, thus giving them identical weights, their *b* parameter estimates should be identical. Table 2 shows considerable differences between Language and Reasoning Materials indicators with identical middle digits, ranging from approximately half an SD to 1 SD difference between these scores (e.g. 16.5.1, 17.5.1, 18.5.1, 18.5.3 and 18.5.4, as well as 16.7.1, 16.7.2 and 18.7.1). Similar trends are evident for indicators in Table 3. Therefore, the *b* parameter estimates for indicators that receive equal weights (i.e. in the present ECERS-R scoring system) vary considerably.

If an indicator is capable of assessing centres with low quality levels but provides limited information for centres with high quality levels, not only does it make little sense to administer this indicator when assessing centres that are higher quality, but totaling indicator scores for centres with different quality levels might lead to misleading scores. Thus, in a system where indicators

Table 3. Parameter estimates for language and reasoning materials.

Brief indicator description	*a*	*b*
15.3.1. Some books	2.82	–0.68
15.3.2. Receptive language	1.51	–1.32
15.5.1. Book selection	4.31	0.08
15.5.2. Additional language	2.53	–0.93
15.5.3. Books organised	2.20	–0.74
15.5.4. Books appropriate	2.10	–1.25
15.7.1. Books rotated	2.17	–0.01
15.7.2. Books relate	2.36	–1.50
16.3.2. Materials accessible	3.78	–0.87
16.5.2. Materials encourage communication	2.37	–0.34

Note: *a* = discrimination parameter estimates; *b* = discrimination parameter estimates.

are totaled or averaged to produce a score on a factor, we can see that we are likely introducing additional 'noise' in the measurement of quality by administering and using indicators that do not reliably measure the quality dimension of interest to be counted in the overall score.

Objective (e): floor and ceiling effects

By addressing whether the subscales are limited by ceiling or floor effects, objective (d) addresses concerns that are similar to objective (e). That is, it addresses whether the subscales are capable of measuring early childhood settings ranging from very low quality to very high quality. Examination of the *b* parameter estimates in Tables 2 and 3 and the test information curves (TICs) for Factors 1 and 2 (shown in Figures 5 and 6) show little by way of floor effects for each factor. In other words, each of these factors include multiple indicators that are all capable of measuring the quality of early childhood environments that are lower than 1 SD below the mean, and, in some cases, almost as low as 2 SDs below the mean. Both scales also have indicators that are capable of measuring quality levels up to and slightly above the mean. Figures 5 and 6, however, show that considered together, the majority of the indicators on both factors are more precise in measuring quality levels at 1 SD below the mean from quality levels that are at the mean. Figure 6 also shows that some of the indicators loading on Language and Reasoning Activities might provide slightly more information on facilities with quality levels that are slightly above the mean than the indicators loading on Language and Reasoning Materials. Nevertheless, both figures show that because the items on both scales show ceiling effects, these scales might not be as precise when measuring settings that possess higher than average quality levels.

Objective (f): appropriateness of how Indicators are ordered

Addressing objective (e) has shown that although the authors of the ECERS-R have calibrated indicators according to face validity, the indicators are incapable of measuring early childhood settings that have high quality levels (i.e. ≥ 0.30 and ≥ .09 SD above the mean for Factors 1 and 2, respectively). Moreover, the results addressing objective (e) show that assigning equal quality level scores for indicators the authors deemed acceptable might be inappropriate. Considered together, these findings suggest that the indicators might not be ordered appropriately. For

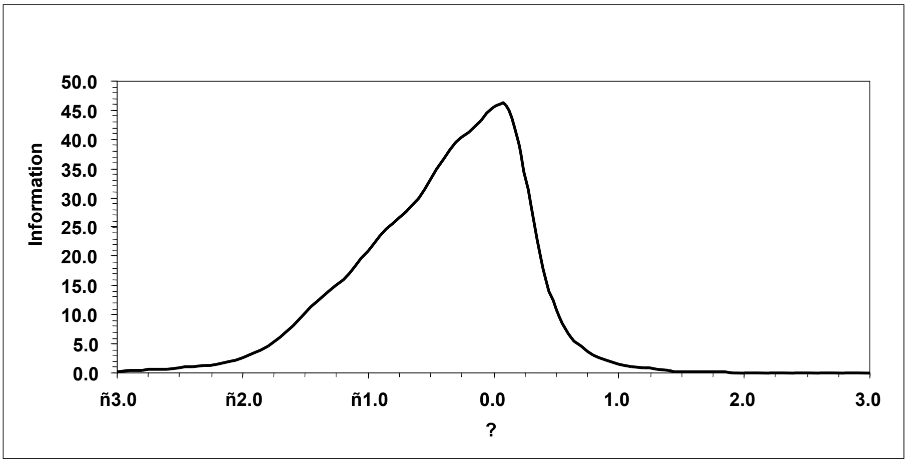

Figure 5. Test information curve for Factor 1.

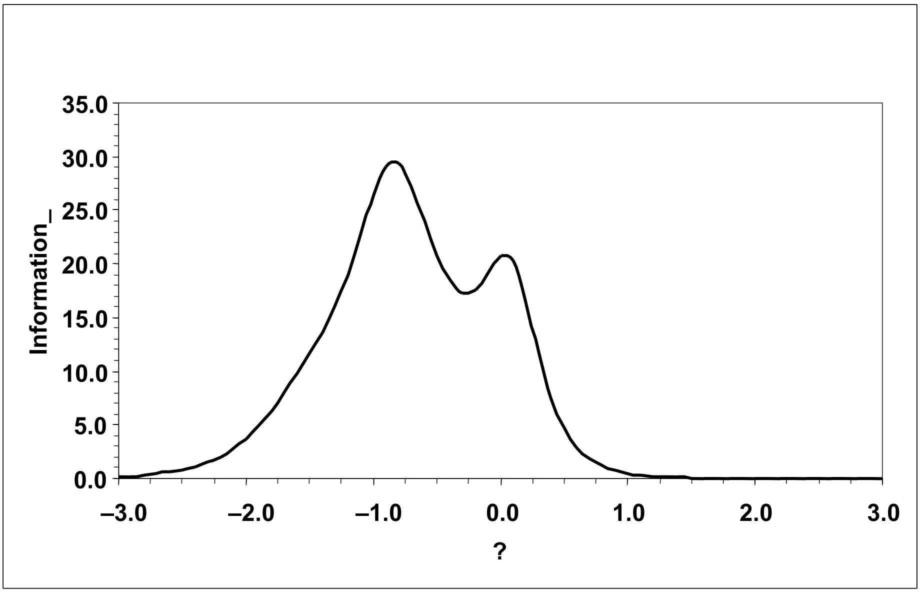

Figure 6. Test information curve for Factor 2.

example, the *b* parameter estimate for 16.3.1 is lower than that for item 16.5.1, but 16.7.1 is lower than 16.5.1 (see Table 2). Similar trends are evident for items loading on Factor 2.

Discussion

The purpose of this study was to use IRT procedures to examine the psychometric properties of the ECERS-R within the context of early childhood settings in the Caribbean. Focusing on the indicators of the language-reasoning scale, one of the preliminary considerations for IRT analyses was to test whether the indicators of this dimension load on a single essentially unidimensional scale. Since they did not satisfy this criterion, we sought to identify factors that were essentially unidimensional. This study also addressed whether indicators loading on their respective factors might provide sufficient information to be used in the assessment of early childhood environment quality, whether it is appropriate to administer all indicators loading on each factor, and whether the practice of equally weighting specific indicators on each factor is appropriate. Finally, it addressed whether the factors possess floor or ceiling effects and whether the authors of the ECERS-R ordered the indicators appropriately.

Dimensionality of the indicators

Given the administration and scoring procedures described in the manual, the authors of the ECERS-R implicitly assume that the indicators that comprise the language-reasoning scale load on a single factor (see Harms, Clifford and Cryer 1998). Findings from the present study show that this assumption might not be plausible, at least within the context of the early childhood settings surveyed in the Caribbean nations of Jamaica and Grenada. Although the two factors derived from IRT-based factor analyses were reflective of language and reasoning, they bifurcated in two distinct factors, one of which was reflective of the educational activities involved in language and reasoning instruction and the other being the materials that might be used in this process. While these factors are related, it is important to note that when their items are considered together they violate the key measurement assumption of unidimensionality and even the more relaxed assumption of essential unidimensionality. Since the indicators undergird the item scores from which ECERS-R scales are derived, this violation might lead to spurious measurement parameter estimates and inaccurate results.

Measurement precision

Having identified the best factor model within the data, another objective was to examine whether the indicators on each factor provided adequate amounts of information to remain as part of the measure. The findings revealed that the indicators on each of the two dimensions are capable of providing appropriate information levels. That is, used in their present form, most indicators appear to be capable of precisely measuring the quality levels their location parameter estimates indicate. However, the indicators on each dimension were identified as being less capable of measuring quality at the higher end of the constructs. From an IRT perspective, these indicators should not be used in cases where they are not appropriate.

Appropriateness of administering and equally weighting all indicators

For each item, the scoring procedures of the ECERS-R require that the early childhood classroom observer should rate indicators as listed under specific Likert scale specific points in order to derive an item score. This procedure not only implicitly assumes that the indicators are ordered correctly, but also that each indicator on each point in the Likert scale is equally weighted. The findings from the present study have revealed that it is inappropriate to equally weight such indicators since they measure very different levels of quality. Moreover, since the quality levels across indicators are often different, it makes little sense to administer all indicators from the respective points of the Likert scale where they are listed, with little regard for the quality of the early childhood setting being measured. It is important that each centre be measured with indicators that match the level of quality it demonstrates. From an IRT perspective, indicators would not be used in cases where they are not appropriate. A useful alternative to totaling indicator scores is to use software that calculates IRT item parameters (e.g. MULTILOG) to estimate an a posteriori score for ratings on all indicators administered within the ECERS-R subscales.

The parameter estimates also show the potential for employing computerised adaptive testing (CAT) – that is, using an item bank (in this case an indicator item bank) with pre-calibrated parameter estimates and matching indicators to the appropriate level of the construct being measured. Such an endeavour might decrease the time and effort required to rate programmes for research purposes. A priori scoring using CAT, where classrooms are rated primarily on items matching their quality levels, could be more economical than a posteriori scoring since theoretically, all items would not be administered.

It might be argued that CAT scoring would be difficult to accomplish since some ECERS-R indicators require the rater’s presence for extended periods of time to appropriately assess the quality levels in facilities being evaluated. While this might be true for some indicators, we do not believe that this is necessarily true for all. For example, some of the indicators measuring higher quality levels on the Language and Reason Activities (e.g. whether staff are talking logically to children, whether children being encouraged to talk), and Language and Reasoning Materials (e.g. wide selection of books being accessible) factors are relatively easy to observe when a rater visits most early childhood settings. The same is true for some of the indicators measuring the lower quality levels, such as some staff–child conversation on Factor 1 and some materials being accessible to help children communicate.

Appropriateness of how ECERS-R indicators are ordered

Findings from the present study show that the indicators for each item might not be appropriately ordered. That is, the *b* parameter estimates for indicators belonging to the two new factors found in this study show that their order, where they best measure quality, do not reflect the authors' order. Findings from assessment and research projects that use item scores that are derived from these indicators might be inaccurate. It is also important to note that by estimating the parameters for indicators from the original language-reasoning scale, scoring items from these indicators as the authors ordered them is now obsolete. That is, since IRT scores are derived from observing the pattern of ratings on each indicator, an early childhood setting will be scored on the basis of this pattern and not on the basis of item scores.

Floor and ceiling effects

The *b* parameter estimates and TICs show that the two dimensions this study identified from the indicators of the language-reasoning scale are capable of measuring quality levels in early childhood settings where quality ranges from well below the mean up to slightly above the mean. The TICs, however, show that, considered together, the items on each dimension are more precise in measuring the quality of early childhood settings at or below the mean. Thus, we suggest that the ECERS-R indicators for the language-reasoning factors are of limited value in assessing the overall quality of the full range of preschool settings, at least in the two Caribbean nations evaluated. They might, therefore, be inappropriate for high-stakes evaluations or any type of tiered reimbursements that might be offered to centres that are moderately or high performing early childhood centres.

Limitations

Compared with some published IRT studies (e.g. Marshall et al. 2002) our sample size is respectable, but we are also aware that one can be more confident regarding item parameter estimates when larger sample sizes are used for such estimation. Moreover, because of the sample size we were unable to examine all indicators within the ECERS-R and therefore restricted our analyses to the indicators that currently measure items under the language-reasoning scale.

Furthermore, we do not know whether the item parameter estimates we calibrated would be psychometrically invariant across raters and across other early childhood settings in the two Caribbean nations surveyed or in other nations across the region. Invariance means that items on a subscale behave in a psychometrically identical fashion across different settings and thus allow unbiased comparisons and aggregation of scores across them. Finally, we recognise that, as is the case for most measures used in social science research, using the ECERS-R to rate the quality of any early childhood setting represents only a snapshot of the quality during the period in which the environment is observed. Ratings derived at the time of assessment might not be fully representative of the environment over time. While one means of addressing this concern might be repeated assessments, measurement artifacts across time might make it difficult to interpret these findings.

Recommendations

Since the indicators from which the language-reasoning items are scored violate the key measurement assumption of unidimensionality or essential unidimensionality, the estimation of the parameters for the indicators has made the use of these 43 ECERS items obsolete. In addition, if the goal for the indicators on the two factors this study identified is merely to screen early childhood environments with substandard to average quality, the indicators comprising the two factors in this study might be appropriately used. If the factors are used to rate early childhood environments that are well above average quality levels, it is critical that indicators reflecting such levels are constructed and evaluated. The built-in linking capabilities of IRT can rigorously and economically facilitate such efforts. While CTT-guided psychometric procedures might be used to construct and evaluate indicators, their sample-dependent nature would most likely require that researchers collect data on all existing ECERS-R indicators in addition to the new indicators in a completely different sample. By collecting data on new indicators and some existing indicators that are invariant across the original and new samples, IRT procedures, by contrast, allow researchers to link both old and new databases and their indicator parameter estimates and thus add indicators that reflect higher quality.

Further research might also test for indicator invariance across sufficiently large numbers of different types of early childhood programmes (e.g. private vs. public settings), across different regions of the Caribbean, and across countries on different continents. Equally important is addressing whether repeated assessment might result in measurement artifacts. IRT is also particularly useful in these endeavours, first because it allows testing for differential item functioning (DIF) across groups and can therefore rule out or identify item bias across different groups of programmes. Second, by examining drift of item parameters (i.e. DIF across different administration points), it can inform the professional as to whether differences in scores are resulting from true differences across assessment points or whether measurement artifacts might be implicated (Meade, Lautenschlager, and Hecht 2005).

Another recommendation is the importance of taking the necessary steps to evaluate how all dichotomously scored indicators in the entire measure might be grouped according to essentially unidimensional factors. Once the indicators are empirically grouped under different factors, calibration of these indicators in a similar manner as was done in this study would also be important. Such calibration would be necessary to continue addressing whether readily observed dichotomous indicators measure quality as well as indicators requiring extensive amounts of time to observe. If the readily observed indicators measure quality levels that are similar to indicators requiring lengthy observation time, and if both sets of indicators discriminate equally across quality levels, the readily observed indicators could then be easily administered by computerised adaptive testing (CAT). With CAT, only items that provide the most information on the specific quality level of each classroom are administered. Such administration would considerably decrease the amount of time needed to measure the quality in a typical early childhood setting for research purposes.

Conclusions

Despite its shortcomings, one of the strengths of this study is that it represents the first effort we know of that has used more modern measurement theory-driven procedures to begin documenting the psychometric properties of the ECERS-R. Moreover, this is the first study that has evaluated whether this measure might be appropriate for use in Caribbean nations such as Jamaica and Grenada. It is our hope that now that we have reported IRT item parameter estimates for the newly derived ECERS-R dimensions of Language and Reasoning Activities and Materials, this information might form a scaffold for further research that addresses some of the concerns raised in this study. This research might include the calibration of dichotomous indicators and possibly further fine-tune the measurement precision of the ECERS-R, especially since it is used across a wide variety of early childhood settings within the Caribbean nations surveyed and across other nations within and outside the Caribbean Basin. We also hope that this study further stimulates the discussion of what constitutes preschool quality in international settings and how it is best measured.

Notes

1. It is important to note that reference to items on the ECERS-R is reflective of 43 items. The scores for each of these items are, however, derived from scores on multiple dichotomously scored indicators (see description of the ECERS-R under the 'Measure' subheading).
2. Our use of indicators is not the typical description used in psychometric theory, but these terms are taken from the ECERS-R and describe the dichotomously scored items from which each of the 43 items is scored.
3. Other methods such as weighted least squares with tetrachoric correlations have been demonstrated to provide similar results to FIFA and can assess data-to-model fit with goodness-of-fit indices (see Woods 2002).
4. TESTFACT allows the researcher to use the marginal maximum likelihood estimate in a bfactor solution that is nested in a previously selected FIFA factor solution and permits the use of a likelihood ratio test for conditional independence. A significantly lower likelihood ratio estimate for a bi-factor solution would suggest the presence of a primary (e.g. second-order) factor. The presence of a second-order factor would provide evidence that conditional independence is questionable and that a second-order factor is necessary to explain the data. In other words, a relationship between indicators would exist even when the trait level is held constant (Panter and Reeve 2002).
5. The likelihood ratio test suggested a three-factor model. That is, the FIFAs where different factor models were compared starting from a one- and a two-factor model revealed that the three- and four-factor models were the first pair of analyses to reveal no significant differences, $\Delta\chi^2$ (25) = 12.92, $p >$ .96. The third-factor in the three-factor solution had loadings accounting for less than 1% of the variance. Furthermore, the three-factor model had multiple cross-loaded indicators, thereby raising concerns regarding the robustness of these factors. The two-factor model was theoretically plausible, since the first factor represented activities that facilitate language and reasoning and the second factor materials used to promote language and reasoning. Procedures Gibbons and Hedeker (1992) used in selecting a factor model based on dominant indicator to factor loadings were therefore used in selecting the two-factor solution used in addressing the next IRT assumption.
6. TESTFACT was used to test for conditional independence by nesting a bi-factor model in the two-factor model. Comparing likelihood ratios across the nested models revealed a nonsignificant effect, $\Delta\chi^2$ (3) = 5.77, p >.10, and showed no significant difference between the bi-factor likelihood ratio and that of a simpler factor structure. This finding suggests that the two-factor model might be appropriately specified and that a primary dimension was not necessary to fully describe the data.
7. For each dimension, the MULTILOG analyses used to test whether 1-, 2-, or 3-parameter models best represented responses were conducted in models where each indicator was unconstrained across the two nations. In the first test, the 1PL model was nested in the 2PL model and in the second test the 2PL model was nested in the 3PL model. For the Language and Reasoning Activities factor the comparisons between the 2PL and the 1PL model showed better fit for the 2PL model, since its G^2 was significantly lower than the G^2 for the1PL model, ΔG^2 (12) = 72.3, $p < .001$. Comparing the 2PL model with the 3PL model showed no better fit for the 2PL model, since its G^2 was not significantly different from the G^2 for the 3PL model, ΔG^2 (108) = 55.8 p >.99, suggesting no significant decrement in the G^2 statistic when 3PL model was applied. The 2PL model was therefore chosen to estimate the parameters for the indicators. Similar trends emerged between comparisons of the 1PL and 2PL models for the Language and Reasoning Materials factor, $\Delta\chi^2$ (9) =29.4,1. $p < .001$. With Δχ2 (10) = 12.4, p >.26 for the 2PL versus the 3PL models. There was no significant difference between the models, suggesting that no significant improvement emerged when the less parsimonious 3PL model was applied.
8. We note that Factors 1 and 2 include 13 and 10 indicators, respectively. Nevertheless, to present uncluttered and legible figures, only four indicators for each factor are selected for illustration.

References

Aboud, F. 2006. Evaluation of an early childhood preschool program in rural Bangladesh. *Early Childhood Research Quarterly* 21: 46–60.

Arnett, J. 1989. Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology* 10: 541–52.

Beller, E.K., M. Stahnke, P. Butz, W. Stahl, and H. Wessels. 1996. Two measures of the quality of group care for infants and toddlers. *European Journal of Psychology of Education* 11, no 2: 151–67.

Bock, R.D., R. Gibbons, S. Schilling, E. Muraki, D. Wilson, and R. Wood, R. 2003. *TESTFACT,* Version 4.0 [Computer software]. Chicago: Scientific Software International.

Burchinal, M., D. Cryer, R. Clifford, and C. Howes. 2002. Caregiving training and classroom quality in child care centers. *Applied Developmental Science* 6: 2–11.

Burchinal, M.R., J.E. Roberts, R. Riggins, Jr., S.A. Zeisel, E. Neebe, and D. Bryant. 2000. Relating quality of center-based child care to early cognitive and language development longitudinally. *Child Development* 71: 339–57.

Calder, P. 1996. Methodological reflections on using the Early Childhood Environment Rating Scale as a measure to make cross-national evaluations of quality. *Early Child Development and Care* 126: 27–37.

Cassidy, D., M. Buell, S. Pugh-Hoese, and S. Russell. 1995. The effect of education on child care teachers' beliefs and classroom quality: Year one evaluation of the TEACH Early Childhood Associate Degree Scholarship Program. *Early Childhood Research Quarterly* 10: 171–83.

Cassidy, D.J., L.L. Hestenes, A. Hedge, S. Hestenes, and S. Mims. 2005. Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the Early Childhood Environment Rating Scale-Revised. *Early Childhood Research Quarterly* 20: 345–60.

Clifford, R., O. Barbarin, F. Chang, D. Early, D. Bryant, C. Howes, M. Burchinal, and R. Pianta. 2005. What is prekindergarten? Characteristics of public pre-kindergarten programs. *Applied Developmental Science* 9: 126–43.

Dunn, L. 1993. Proximal and distal features of day care quality and children's development. *Early Childhood Research Quarterly* 8: 167–92.

Farquhar, S. 1989. Assessing New Zealand child day care quality using the early childhood environment rating scale (1). *Early Child Development and Care* 47: 93–105.

Gibbons, R. D. & D. R. Hedeker. 1992. Full information item bi-factor analysis. *Psychometrika* 57: 423–435.

Goelman, H., B. Foere, P. Kershaw, G. Doherty, D. Lero, and A. LaGrange. 2006. *Toward a predictive model of quality in Canadian child care centers* 21: 280–95.

Hambleton, R. K., H. Swaminathan, & H. J. Rogers, 1991. *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Harms, T., and R.M. Clifford. 1980. *Early childhood environmental rating scale.* New York: Teachers College Press.

Harms, T., R. Clifford, and D. Cryer. 1998. *Early Childhood Environment Rating Scale Revised Edition.* New York and London: Teachers College Press.

Hattie, J. 1996. An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement* 20: 1–14.

Hattie, J., K. Karkowski, & H. J. Rogers 1996. An assessment of Stout's Index of essential unidimensionality. *Applied Psychological Measurement* 20: 1–4.

Helburn, S. 1995. *Cost, quality, and child outcomes in child care centers.* Available from the Department of Economics, Center for Research in Economic and Social Policy, University of Colorado at Denver.

Herrera, M.O., M.E. Mathiesen, J. Merino, and I. Recart. 2005. Learning contexts for young children in Chile: Process Quality Assessment in Preschool Centres. *International Journal of Early Years Education* 13: 13–27.

Holloway, S., S. Kagan, B. Fuller, L. Tsou, and J. Carroll. 2001. Assessing child-care quality with a telephone interview. *Early Childhood Research Quarterly* 16: 165–89.

Karrby, G., and J. Giota. 1994. Dimensions of quality in Swedish day care centers – an analysis of the early childhood environment rating scale. *Early Child Development and Care* 104: 1–22.

Lee, Y.-J., J.-S. Lee, and J.-W. Lee. 1997. The role of the play environment in young children's language development. *Early Child Development and Care* 139: 49–71.

Marshall, G.N., M. Orlando, D.W. Foy, and H. Blezberg. 2002. Development and validation of a modified version of the Peritraumatic Dissociative Questionnaire. *Psychological Assessment* 14: 123–34.

McCarty, F., Abbott-Shim, R. Lambert. 2001. The relationship between teacher beliefs and practices, and Head Start classroom quality. *Early Education and Development* 12: 225–238

Meade, A. W., G. J. Lautenschlager, J. E. Hecht. 2005. Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing* 5: 279–300.

Moore, K., T. Halle, S. Vandivere, and C. Mariner. 2002. Scaling back survey scales: How short is too short? *Sociological Methods and Research* 30: 530–66.

Nandakumar, R. 1993. Assessing essential unidimensionality of real data. *Applied Psychological Measurement* 17: 29–38.

National Institute of Child Development Early Child Care Research Network. 2000. The relations of early child care to cognitive and language development. *Child Development* 71: 960–80.

Panter, A.T., and B.B. Reeve. 2002. Assessing tobacco beliefs among youth with item response theory models. *Drug and Alcohol Dependence* 68: 21–39.

Peisner-Feinberg, E., M. Burchinal, R. Clifford, M. Culkin, C. Howes, S. Kagan, N. Yazejian. 2001. The relation of preschool child-care quality to children's cognitive and social development trajectories through second grade. *Child Development* 75: 1534–53.

Perlman, M., G. Zellman, and V. Lee. 2004. Examining the psychometric properties of the early childhood environment scale-revised (ECERS-R). *Early Childhood Research Quarterly* 19: 398–412.

Phillips, D., D. Mekos, S. Scarr, K. McCartney, and M. Abbott-Shim. 2000. Within and beyond the classroom door: Assessing quality in child care. *Early Childhood Research Quarterly* 15: 475–96.

Phillipsen, L.C., M.R. Burchinal, C. Howes, and D. Cryer. 1997. The prediction of process quality from structural features of child care. *Early Childhood Research Quarterly* 12: 281–303.

Sakai, L., M. Whitebook, A. Wishard, and C. Howes. 2003. Evaluating the Early Childhood Rating Scale (ECERS): Assessing differences between the first and revised edition. *Early Childhood Research Quarterly* 18: 427–45.

Scarr, S., M. Eisenberg, and K. Deater-Decker. 1994. Measurement of quality in child care centers. *Early Childhood Research Quarterly* 9: 131–51.

Sheridan, S. 2001. Quality evaluation and quality enhancement in preschool: A model of competence development. *Early Child Development and Care* 166: 7–27.

Stout, W.F. 1990. A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika* 55: 293–325.

———. 2005. POLY-DIMTEST manual. Champaign, IL: The William Stout Institute for Measurement.

Stout, W.F., B. Habing, J. Douglas, H.R. Kim, L. Roussos, and J. Zang. 1996. Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement* 20: 331–54.

Thissen D., W. Chen, & R. D., Bock. 2003. MULTILOG (2003). Lincolnwood, IL. Scientific Software International.

Thompson, T.D., and M. Pommerich. 1996. Examining the effects of local dependence. Paper presented at the annual meeting of the American Educational Research Association, April, in New York.

Tietze, W., D. Cryer, J. Bairrao, J. Palacios, and G. Wetzel. 1996. Comparisons of observed process quality in early child care and education programs in five countries. *Early Childhood Research Quarterly* 11: 447–75.

Vandell, D. L., B. Wolfe. 2000. *Child care quality: does it matter and does it need to be improved?* Wisconsin: Wisconsin Univ., Madison. Inst. for Research on Poverty.

Warash, B., C. Markstrom, and B. Lucci. 2005. The early childhood environment rating scale-revised as a tool to improve child care centers. *Education* 126: 240–50.

Wilson, D. T, R. D. Bock, R. D. Gibbons, S. Schilling, E. Muraki & R. Wood. 2003. TESTFACT. Lincolnwood, IL. Scientific Software International.

Woods, C. M. 2002. Factor analysis of scales composed of binary items: Illustration with the Maudsly Obsessional compulsive scale. *Journal of Psychological and Behavioral Assessment* 24: 215–223.