Psychometric Properties and Differential Item Functioning of a Web-Based Assessment of Children's Emotion Recognition Skill

Author: Beyza Aksu Dunya (ORCID: 0000-0003-4994-1429), Clark McKown (ORCID: 0000-0001-9694-1179), Everett Smith

Psychometric Properties and Differential Item Functioning of a Web-Based Assessment of Children's Emotion Recognition Skill

Beyza Aksu Dunya[1,2], Clark McKown[3] iD**,
and Everett Smith[2]**

Abstract

Emotion recognition (ER) involves understanding what others are feeling by interpreting nonverbal behavior, including facial expressions. The purpose of this study is to evaluate the psychometric properties of a web-based social ER assessment designed for children in kindergarten through third grade. Data were collected from two separate samples of children. The first sample included 3,224 children and the second sample included 4,419 children. Data were calibrated using Rasch dichotomous model. Differential item and test functioning were also evaluated across gender and ethnicity. Across both samples, we found consistent item fit, unidimensional item structure, and adequate item targeting. Analyses of differential item functioning (DIF) found six out of 111 items displaying DIF across gender and no items demonstrating DIF across ethnicity. The analyses of person measure calibrations with and without DIF items yielded no evidence of differential test functioning (DTF) across gender and ethnicity groups in both samples.

Many social and emotional competencies influence children's ability to succeed in relationships, in school, and in life (McKown, 2017). Some of those competencies are thinking skills involved in understanding others' emotions and intentions, solving social problems, and regulating emotions. Collectively, we refer to these thinking skills as "social-emotional comprehension" (Lipton & Nowincki, 2009). The better developed is children's social-emotional comprehension, the better they do in a range of functional outcomes (Banerjee & Watling, 2005; Blair & Razza, 2007; Denham, 2006; Nowicki & Duke, 1994; McKown et al, 2016). Furthermore, social and

[1]Bartin University, Turkey
[2]University of Illinois at Chicago, USA
[3]Rush University Medical Center, Chicago, IL, USA

Corresponding Author:
Clark McKown, Rush Neurobehavioral Center, Department of Behavioral Sciences, Rush University Medical Center, 4711 Golf Road, Suite 1100, Skokie, IL 60076, USA.
Email: Clark_A_McKown@rush.edu

emotional competencies are increasingly the focus of state social and emotional learning standards (Dusenbury, Dermody, & Weissberg, 2018) and universal and indicated instructional programs (Weissberg, Goren, Domitrovic, & Dusenbury, 2012).

Despite its importance and growing integration with educational practice, few tools are available for educators and other professionals to assess children's social-emotional comprehension. As a result, professionals charged with teaching these important skills are often left without information about student strengths and needs—information they can use to guide how and what they teach to whom and to measure student skill acquisition. Rigorous assessments designed to measure children's social-emotional comprehension can therefore support educators in their efforts to nurture children's social and emotional competence.

As noted elsewhere (McKown, Russo-Ponsaran, Allen, Johnson, & Russo, 2016), we believe optimal assessments will (a) adequately sample the content domain (Nunnally & Bernstein, 1994), (b) be easy for educators and other professionals to use, (c) permit group administration (Murphy & Davidshofer, 2004), and (d) be appropriate for a broad population of children. Finally, because social-emotional learning is a priority in early elementary school (Thompson & Goodman, 2009), its assessment is particularly important in the early grades.

To address the need for social-emotional comprehension assessments with these characteristics, we developed a web-based assessment called SELweb. SELweb includes five different subtests, or "modules," designed to assess four distinct but interrelated dimensions of social-emotional comprehension, including one module each to assess children's understanding of facial expressions (emotion recognition [ER]), children's ability to infer another person's intentions or beliefs (social perspective-taking), children's understanding of and ability to solve social problems (social problem-solving), and two modules to assess children's ability to voluntarily modulate thoughts and feelings (self-control).

SELweb is administered by a web application and can be group administered in schools. It assesses skills that are the focus of instruction in many commonly used social and emotional learning programs. As a result, educators can use what they from SELweb to guide instructional decision-making. For example, if a teacher learns that children struggled to recognize others' emotions, that teacher might opt to emphasize lessons that focus on this skill, which are often incorporated in evidence-based SEL curricula (Weissberg et al., 2012). SELweb is also a cost-effective assessment for researchers whose work focuses on children's social and emotional competencies. It is currently being used as an outcome measure in several field trials of social and emotional programs and interventions.

We conceptualize each of the four dimensions of social-emotional comprehension SELweb assesses as a partially independent component of social-emotional comprehension. That conceptualization is supported by two empirical studies showing an excellent fit of the data to a four-factor confirmatory model in which ER, social perspective-taking, social problem-solving, and self-control are modeled as correlated latent variables (McKown et al, 2016; McKown, 2019). Together, SELweb's modules provide broad construct coverage of distinct but interrelated dimensions of social emotional comprehension. ER, the focus of this article, is one of the four factors in that model.

Previous research examining all of SELweb modules together has described its psychometric properties, including evidence of structural validity, and discriminant, convergent, and criterion-related validity (McKown et al., 2016). In addition, using a confirmatory factor analysis approach, (McKown, 2019) reported that SELweb's five modules demonstrated configural, metric, and partial scalar invariance across sex and ethnicity. The confirmatory models in that measurement equivalence study used total scores from the five modules as indicators of latent constructs. As a result, those findings provide information about the comparability of total scores for children from different groups, which reflects differential test functioning (DTF).

The validity evidence cited above reflected a classical test theory approach focused on evaluating the technical properties of total scores. An alternative to validation involves an approach described by Rasch (1960) focused on examining the underlying properties of the items that make up those scores. The Rasch approach complements and extends prior work by providing a focused analysis of the extent to which items and the totals to which they contribute reflect several forms of validity.

In addition, a Rasch approach provides an alternative and complementary framework for evaluating measurement equivalence. Evidence of DTF as described above does not answer the question of whether items that make up the total scores function similarly for children from different groups, which is a question of differential item functioning (DIF). DIF analysis focuses on items within a scale designed to measure a single construct. The focus on items measuring a single construct, referred to as "unidimensionality," is a key assumption of DIF (Rasch, 1960). Therefore, to evaluate DIF in SELweb requires analyses that focus separately on items within each module, because those items are designed to measure a single dimension of social-emotional comprehension.

For example, one paper evaluated DIF in SELweb's social perspective-taking module (McKown et al, 2016), one of the four constructs SELweb is designed to assess. Those analyses supported the unidimensionality of items in this module and revealed negligible DIF. Building on that work, the purpose of the present study is to evaluate the psychometric properties of SELweb's ER module using a Rasch (1960) analytic framework and separately examining two diverse and independent samples ($N_1 = 4,464$, $N_2 = 3,220$) that included a total of 7,643 children. A particular focus of this study is to evaluate DIF across gender and ethnicity.

Assessing ER

Rationale

We included an ER assessment in SELweb because emotions play a key role in human social interactions. Defined as the ability to read nonverbal cues that signal what others feel, ER is associated with a range of functional domains, including internal locus of control, self-esteem, and peer acceptance and that children's ER skill was related to reading and math achievement (Nowicki & Duke, 1994). Typically, ER assessments involve viewing photographs of people's faces and indicating what the person is feeling from their facial expression (Nowicki & Duke, 1994; Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979).

Existing ER Assessments

Existing ER assessments have strengths, but none has all four of the desirable assessment characteristics described previously. Many have been used for research purposes to characterize the social impairments in clinical populations. For example, Tehrani-Doost et al. (2017) used the Facial Emotion Recognition Task (FERT) to compare 7- to 12-year-old boys who have been diagnosed attention-deficit/hyperactivity disorder (ADHD) with typically developing children. The FERT assesses recognition of male and female facial expressions with the same intensity. They found that children with ADHD were less sensitive to both positive and negative emotions. Another study by Wyssen et al. (2019) used the Difficulties with Emotion Regulation Scale (DERS), developed by Gratz and Roemer (2004), to compare recognition of negative emotions among women experiencing eating disorders with ER among healthy women and women experiencing mood and anxiety disorders. These studies contribute to our understanding of social cognitive challenges in clinical populations. However, it is unclear that

Table 1. Widely Used and Existing Emotion Recognition Assessments.

Assessment	Reference	Description
SELweb Emotion Recognition	McKown, Russo-Ponsaran, Johnson, Russo, and Allen (2016)	SELweb is a nationally normed, web-based assessment for kindergarten to third grade designed to assess emotion recognition, social perspective-taking, social problem-solving, and self-control.
DANVA	Nowicki and Duke (1994)	The DANVA measures recognition of child and adult facial expression, tone of voice, and posture at high and low intensities. It is computer delivered and has provisional norms.
UCDSEE	Tracy and Robins (2004)	The UCDSEE measures recognition of adult facial expressions and postures, including social emotions such as pride, embarrassment, and guilt. It is not normed.
NEPSY Affect Recognition	Korkman, Kirk, and Kemp (2007)	The NEPSY is a neuropsychological assessment battery for young children that includes a facial emotion recognition module. It is administered and scored by a trained test administrator.
POFA	Ekman and Friesen (1976)	The POFA includes 110 photographs of adult faces in one of six emotion displays. The POFA faces have mostly been used for research purposes.
MSCEIT-YV	Mayer, Salovey, and Caruso (2006)	The MSCEIT-YV includes multiple modules assessing several dimensions of emotional intelligence, including perceiving emotions, which is equivalent to emotion recognition.

Note. DANVA = diagnostic analysis of nonverbal accuracy; UCDSEE = UC Davis Set of Emotion Expression; NEPSY = A Developmental NEuroPSYchological Assessment; POFA = Pictures of Facial Affect; MSCEIT-YV = Mayer-Salovey-Caruso Emotional Intelligence Test, Youth Version.

the measures they use are appropriate for the wide-scale assessment of typically developing children to inform instruction.

Table 1 lists some widely used and available ER assessments and their characteristics. For example, most current ER assessments, such as the UC Davis Set of Emotion Expression (Tracy & Robins, 2004), the NEPSY Affect Recognition test (Korkman, Kirk, & Kemp, 2007), the Pictures of Facial Affect (POFA; Ekman & Friesen, 1976), and the Mayer-Salovey-Caruso Emotional Intelligence Test, Youth Version (MSCEIT-YV; Mayer, Salovey, & Caruso, 2004), do not vary the intensity of the facial expressions depicted. However, in reality, people express emotions at different levels of intensity, and sensitivity to these emotional signals is an important feature of ER. Although a small number of studies have examined facial affect recognition at different affect display intensities (Herba, Landau, Russell, Ecker, & Phillips, 2006; Montirosso, Peverelli, Frigerio, Crespi, & Borgatti, 2010; Nowicki & Duke, 1994), no instruments have been developed to assess ER across a large range of item difficulties. Most of the existing ER assessments have been validated using relatively small and homogeneous samples in terms of gender, ethnicity, and cultural backgrounds. None of these assessments is feasible to group-administer in schools. In terms of broad construct representation, two of these assessments—the NEPSY and

Table 2. Sample Descriptive Statistics.

Characteristic	Sample 1		Sample 2	
	n	(%)	*n*	(%)
Gender				
Female	2,230	(50.0)	1,581	(50.9)
Male	2,234	(50.0)	1,639	(49.1)
Ethnicity				
White	1,972	(44.2)	1,828	(56.8)
Black	575	(12.9)	132	(4.1)
Hispanic	1,409	(31.6)	873	(27.1)
Other	209	(4.7)	455	(5.2)
Grade				
K	780	(17.5)	494	(15.3)
1	1,257	(28.2)	985	(30.6)
2	1,360	(30.5)	889	(27.6)
3	1,067	(23.9)	852	(26.5)

the MSCEIT-YV—sample dimensions of social-emotional comprehension other than ER. Finally, we are aware of no evidence of the measurement equivalence, either DTF or DIF, of any of these assessments, and therefore it is not possible to determine whether they function equally well for different subgroups of learners.

SELweb's ER Assessment

SELweb's ER assessment includes pictures of school-aged children's faces, with emotion expressions that range in intensity from very subtle to very strong. Prior research has found that SELweb's ER assessment exhibits good internal consistency reliability ($\alpha \approx .85$), is correlated with, but distinct from, other dimensions of social-emotional skill, and, along with those other dimensions, is positively associated with important outcomes such as socially competent behavior, academic skills, and social acceptance (McKown et al., 2016). SELweb is web-based and can be administered to groups. It therefore is designed to have the characteristics needed to be useful to educators—it samples social and emotional domains widely, is easy to use, and can be group administered at large scale. One important question is whether and to what extent SELweb's ER module works similarly for children from diverse backgrounds. The present study applies Rasch analysis to SELweb's ER assessment to evaluate SELweb ER's psychometric properties, including dimensionality, item fit, and DIF by gender and ethnicity.

Method

Participants and Procedures

Data were collected from two large and diverse samples of students in general education classrooms. As summarized in Table 2, the first sample included 4,464 children and the second sample included a total of 3,220 children. Both samples spanned kindergarten through third grade.

Instrument

Six photographs of child faces with neutral facial expressions, including three girls and two ethnic minorities, were used to create the ER assessment. Children depicted in the images were in

first through fourth grades. The photographs were digitized using FaceGen software (Singular Inversions, 2005). They were then and altered into high-intensity displays of happy, sad, angry, and frightened. To confirm that each face communicated the intended emotion, high-intensity emotion displays were coded by a consultant trained in the Facial Affect Coding System (FACS; Ekman, Friesen, & Hager, 2002), which is an objective coding system used to characterize facial expressions. Faces were iteratively revised until all faces clearly and distinctly displayed its intended emotion.

For each of the six faces and four emotions, we created a set of 10 faces ranging from low- to high-intensity affect displays, forming a pool of 240 images or items. From this item pool, five different test forms were created with 40 items each. Faces were assigned to test forms to ensure a balance of emotions, intensities, and child faces within a given form. Sixteen to 20 items on each test form were included on more than one form. Common items across forms permitted the forms to be linked through a single Rasch analysis. The total number of images used in the assessment was 111.

After each face was presented, children clicked to indicate whether the face reflected happy, sad, angry, scared, or just okay. Use of a web-based format supported the assessment's feasibility when applying, recording, and scoring. School personnel did not need to record or code children's responses. For this article, responses were scored dichotomously, with correct responses awarded 1 point and incorrect responses assigned 0 points. A correct response scored as 1-point meant identification of accurate emotion and intensity by a child.

Analyses

For both samples, item and person data were calibrated using Rasch (1960) dichotomous model. We elected to use Rasch model because it overcomes many limitations of true score theory such as the sample dependency of item and test indices and the item dependency of person's ability. In Rasch measurement, when the data fit model expectations, the item and person parameters are freed from the distributional properties of incidental parameters.

With Rasch model, the probability of giving an answer is quantified as a function of person and item parameters:

$$\Pr\{X_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}}, \tag{1}$$

where Pr is the probability of examinee *n* scoring 1 on item *i*, δ is the difficulty parameter of item *i*, and β is the ability parameter of examinee *n*.

Analyses were performed with WINSTEPS, version 3.93.0 (Linacre, 2019). We employed Wolfe and Smith's (2007a, 2007b) interpretation of Messick's (1995) validity framework along with views articulated by the Medical Outcomes Trust (MOT; see http://www.outcomes-trust. org) to evaluate the psychometric properties of the ER assessment. Under Messick's and the MOT framework, we addressed the content, substantive, structural, responsiveness, and generalizability aspects of construct validity. Each aspect of validity and its associated evaluation criteria is described below.

Content validity. Content validity refers to the representativeness and technical quality of the items (Messick, 1995; Wolfe & Smith, 2007b). To evaluate content validity, we seek answer to the question, "Do the items in the measurement instrument address the intended latent variable?" This question can be addressed using expert judgments, documentation of the instrument development process, and appropriate item fit indices (E. V. Smith, 2002; Wolfe & Smith, 2007b). In Rasch analysis, when assessing items' technical quality via fit indices, we examine point-measure correlations and item fit

statistics. The point-measure correlation (analogous to the traditional item-total correlations expect calculated with Rasch measures) ranges from −1 to +1, with values of .4 or better are preferred, as the size of the correlation indicates that item-level observed scoring accords with the latent variable (Linacre, 2008). As the indication of item fit, Outfit mean-square and Infit mean-square indices are calculated. Outfit mean-square quantifies the degree to which item responses adhere to Rasch model expectations. It indicates technical quality of an item and is sensitive to unexpected observations by persons on items that are too easy or difficult for them (Linacre, 2008). Its expected value is 1.00 and values that are greater than 2.00 distorts the measurement system and inferences made from scores (Linacre, 2008). Unlike Outfit, Infit is more sensitive to unexpected patterns of responses by persons on items that are targeted to their ability level (Linacre, 2008). In Infit, each squared standardized residual is weighted by the information function, so Infit is less influenced by outliers. In the Rasch context, Outfit is generally preferred to Infit unless there is a strong reason to proceed otherwise, such as when data are heavily contaminated with irrelevant outliers (Linacre, 2008). Furthermore, based on simulation studies, the Outfit statistic generally has more power than Infit in detecting measurement disturbances (R. M. Smith, 1991). Therefore, we used the Outfit mean-square item fit statistics to evaluate the fit of the data to model expectations. As suggested by R. M. Smith, Schumacker, and Bush (1998), we flagged an item when the associated Outfit mean-square statistic value was above 2.00. We also checked the point-measure correlation values as additional evidence of item fit (Wolfe & Smith, 2007b). Linacre (2008) suggested flagging any item with a point-measure correlation value that is less than .40.

Substantive validity. Substantive validity refers to "theoretical rationales for the observed consistencies in test responses" (Messick, 1995, p. 745). In the Rasch framework, substantive aspect of construct validity can be assessed by analyzing a WINSTEPS-produced variable map, called Wright item-person map. This map shows distribution of persons and items vertically, with the highest performing persons and the hardest items at the top. Along with the variable map, substantive validity was evaluated with person fit statistics and comparisons of the empirical with the theorized item difficulty hierarchy. Similar to the interpretation of item fit statistics, person fit statistics address the adherence of a person's observed response compared with those predicted by the model. Therefore, Outfit mean-square person fit indices were checked to see whether every child responded to the item difficulty hierarchy as expected. Based on the simulation results by R. M. Smith et al. (1998), children with Outfit mean-square person fit indices larger than 2.00 were flagged as exhibiting a response pattern that does not fit the expected responses based on the item hierarchy.

Structural validity. When developing a measurement instrument, it is expected that theory of the construct domain guides construction and selection of items (Messick, 1995). Validation efforts include ensuring this principle by addressing structural validity. Structural validity indicates the degree to which the structure of the scored observations conforms to the construct domain (Messick, 1995). First, we checked whether the response data reflect a single underlying construct. Principal components analysis (PCA) of residuals identifies patterns in the data that do not accord with underlying construct and is therefore a commonly used method for analyzing any potential pattern of in the data that may reflect secondary dimensions that may not be captured by the item fit statistics (R. M. Smith, 2002). In the Rasch context, a secondary dimension needs to have the strength of at least two items (Linacre, 2008). The strength of a dimension is quantified by *eigenvalues*, which are produced by WINSTEPS as part of PCA of residuals output. To obtain baseline values for comparing eigenvalues, we simulated Rasch-fitting, unidimensional data in WINSTEPS and compared PCA results of the simulated data with the empirical results from Samples 1 and 2. Eigenvalues less than 2.00 imply that the contrast occurred as random noise rather than implying a secondary dimension (Linacre, 2003). An eigenvalue greater than 2.00 may imply a

systematic pattern (underlying construct) in the residuals. Stevens's (2002) criteria provide a framework for interpreting the meaning of eigenvalues greater than 2.00. Specifically, a contrast with eigenvalue greater than 2.00 may be a separate dimension if (a) at least three items with absolute loadings greater than 0.80 were loaded on it, (b) at least four items with absolute loadings greater than 0.60 were loaded on it, and (c) at least 10 items with absolute loadings greater than 0.40 were loaded on it (Stevens, 2002).

Responsiveness. Responsiveness means an assessment tool's capacity to detect change (R. M. Smith, 2002; Wolfe & Smith, 2007b). It can also be conceived as number of statistically distinct level of person measures that can be distinguished by the assessment items (R. M. Smith, 2002). Rasch person separation indices are analyzed as measures of responsiveness. The criteria for evaluating responsiveness was that low person separation ($<$2, equivalent to person reliability $<$.80) with a relevant sample implies that the assessment may not be sensitive enough to distinguish between high and low levels of the construct being measured. Thus, we expected a person separation value greater than 2.00 to ensure the ER items' responsiveness.

Generalizability. Finally, we evaluated generalizability, which refers to the invariance of inferences made from parameter estimates across different groups (i.e., gender, grade level), time points, and contexts by conducting DIF analyses. DIF is routinely assessed in instrument development and validation to ensure measurement invariance, fairness, and generalizability of results across groups. We employed the equal mean differences (EMD) approach (Wang, 2004), a widely used approach for testing DIF between two groups, for gender DIF and WINSTEPS between-group fit statistics, which is appropriate when assessing three or more groups, for testing ethnicity DIF. In the EMD approach, two separate calibrations were run in WINSTEPS to obtain two sets of item parameters, one set for each group. In this approach, the difference between item parameter estimates from separate groups can be directly compared. The size and significance of the DIF effect are examined by employing the separate calibrations z-test approach (R. M. Smith, 2004) by calculating z statistics for each item in both samples:

$$z = \frac{d_{i1} - d_{i2}}{\sqrt{\left(s_{i1}^2 + s_{i2}^2\right)}},$$

where d_{i1} and d_{12} reflect item difficulties based on subgroups and s_{i1} and s_{i2} reflect standard errors for item difficulties.

When DIF items were observed, we also examined the significance of DIF at the test-level, which is referred to as DTF. For testing DTF, we calculated two sets of person measure estimates, one set obtained from a model in which DIF items were treated as DIF-free and one set obtained from a model where DIF items excluded (Wang, 2004). Then, we correlated person measure estimates obtained from these two sets of calibrations. The graphical evaluation of DIF between gender groups at test-level was conducted via test characteristic curve (TCC), obtained for girls and boys separately. The TCC is a cumulative distribution plot of expected test score against ability. When item difficulty invariance holds across groups, TCC remains the same regardless of the ability distribution of the group.

DIF across ethnicity groups was examined using between-group fit statistics (R. M. Smith & Plackner, 2009) produced by WINSTEPS. Between-group fit statistics have an expected value of 1.0. A value that is greater than 1.0 indicates a divergence from model expectations and a value that is smaller than 1.0 indicates overfit. A t statistic, t = ZSTD, associated with between-group fit statistics, was computed for each comparison.

Results

Content Validity Evidence

Looking at the item fit results, only one item showed a slight misfit with an Outfit mean-square value of 2.11 in the first sample. This item is among the difficult items with a difficulty measure of 3.12. None of the item's Outfit mean-square value exceeded 2.00 in the second sample. The point-measure correlation values were all above .40 which indicates consistently robust associations between item measures and the average of all other items. The composite results from both samples supported evidence of content validity for the ER scale.

Substantive Validity Evidence

Figure 1(a) and (b) shows Wright item-person maps for Samples 1 and 2, respectively. On each figure, the left-hand column locates person ability measures' spread along the latent variable (represented by #) and right-hand column locates item difficulty measures (represented by X). Looking at the figures, it can be concluded that items are well targeted to the children for both samples. The mean item difficulty and mean person ability matched for Sample 1, with few extreme children, denoted by "." at both ends of the figure. The mean item difficulty was slightly below the mean person ability with fewer extreme persons in Sample 2, compared with Sample 1. Despite negligible number of extreme persons, both samples yielded similar results in terms of targeting of items to the children. For most of the children, except few with very low and very high ability, measures were targeted efficiently by ER items. Examining the person fit statistics, a total of 27 (0.6%) children showed evidence of significant misfit in the first sample. To further investigate persons with abnormal response patterns, we examined the standardized residuals between observed response and the response according to the Rasch Model. The residual analysis revealed that most of the person misfit occurred as a result of unexpected correct responses to some of the happiness items by a group of overall low-performing children. A similar pattern was observed in the second sample. Out of 3,220 children, only 29 (0.9%) showed significant misfit. Inspection of residuals again suggests unexpected correct responses to some happiness items. Both samples yielded very small percentage of misfitting children with a small percentage of unexpected responses to a subgroup of happiness items. The item difficulty hierarchies displayed a consistent pattern across empirical results from Sample 1 and Sample 2. The findings from multiple analyses supported the substantive validity evidence of the ER scale.

Structural Validity Evidence

PCA of the residuals revealed that the Rasch dimension explained 27.9% of the variance in the first sample. In simulated data, the Rasch component accounted for 41.7% of the variance. The first contrast in residuals explained 3.3% of the variance. The eigenvalue of the first contrast was 3.7, which is larger than the minimum to consider this a dimension (Linacre, 2002). In the second sample, the Rasch dimension explained 38.7% of the variance in the data, whereas the Rasch dimension explained 42.5% of the variance in the Rasch-fitting simulated data. Similarly, in the second sample, the first contrast in residuals accounted for 3.3% of the variance with an associated eigenvalue of 3.6. However, based on Stevens's (2002) criteria, described previously, the results suggested that the residual components did not warrant further interpretation, as only two of the items had loadings greater than 0.40.

Evidence for Responsiveness

As an indicator of responsiveness, person separation index was 2.18 with a person reliability of .88 for Sample 1. Similarly, the person separation index was 2.21 with a person reliability of .84
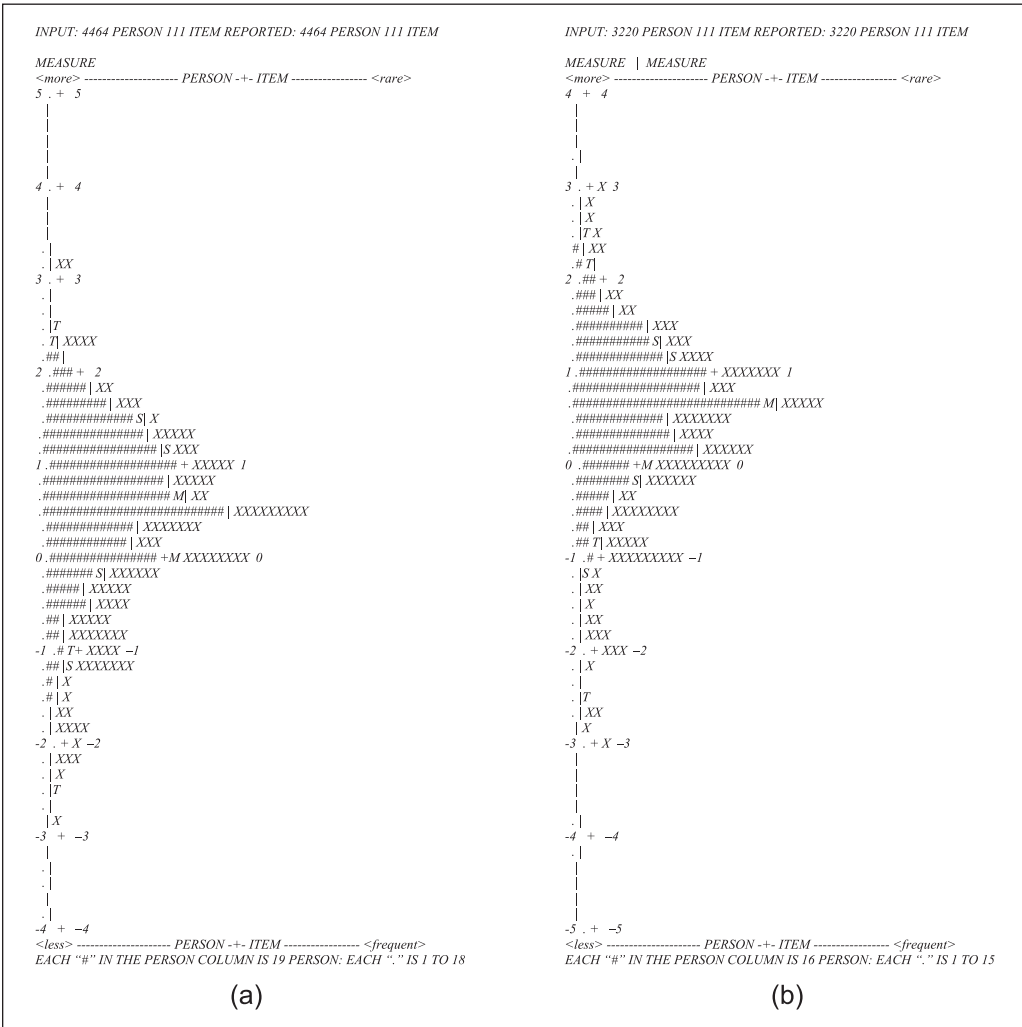
Figure 1. Variable map for (a) Sample 1 and (b) Sample 2.

for Sample 2. The person separation index values obtained from both samples were above the minimum expected value of 2.00. The person reliability values associated with separation indices were above .80 which indicates the instruments' capacity to detect different strata of children's performance.

Generalizability Evidence

DIF analyses revealed that six out of 111 items showed evidence of significant DIF between gender groups on both samples. A closer look to the DIF items from both samples revealed that the same set of four items measuring happiness recognition favored girls and two items measuring anger recognition favored boys (see Table 3). To assess the practical importance of those DIF items on test functioning, we calibrated person measures with and without the DIF items and then correlated the two sets of person measure estimates. The Pearson correlation coefficient between person measures with and without DIF items was .99 ($p < .005$) in the first sample and .99 ($p <$

Table 3. Summary of the Items That Showed DIF Between Gender Groups.

Item name and represented emotion	Sample 1		Sample 2	
	z	Interpretation of DIF	z	Interpretation of DIF
H7I (Happiness)	−3.01	Easier for girls	−3.57	Easier for girls
H8H (Happiness)	−2.74	Easier for girls	−3.05	Easier for girls
H6I (Happiness)	−2.61	Easier for girls	−2.89	Easier for girls
H9I (Happiness)	−2.54	Easier for girls	−2.65	Easier for girls
A2L (Anger)	2.92	Easier for boys	3.12	Easier for boys
A6A (Anger)	2.56	Easier for boys	2.65	Easier for boys

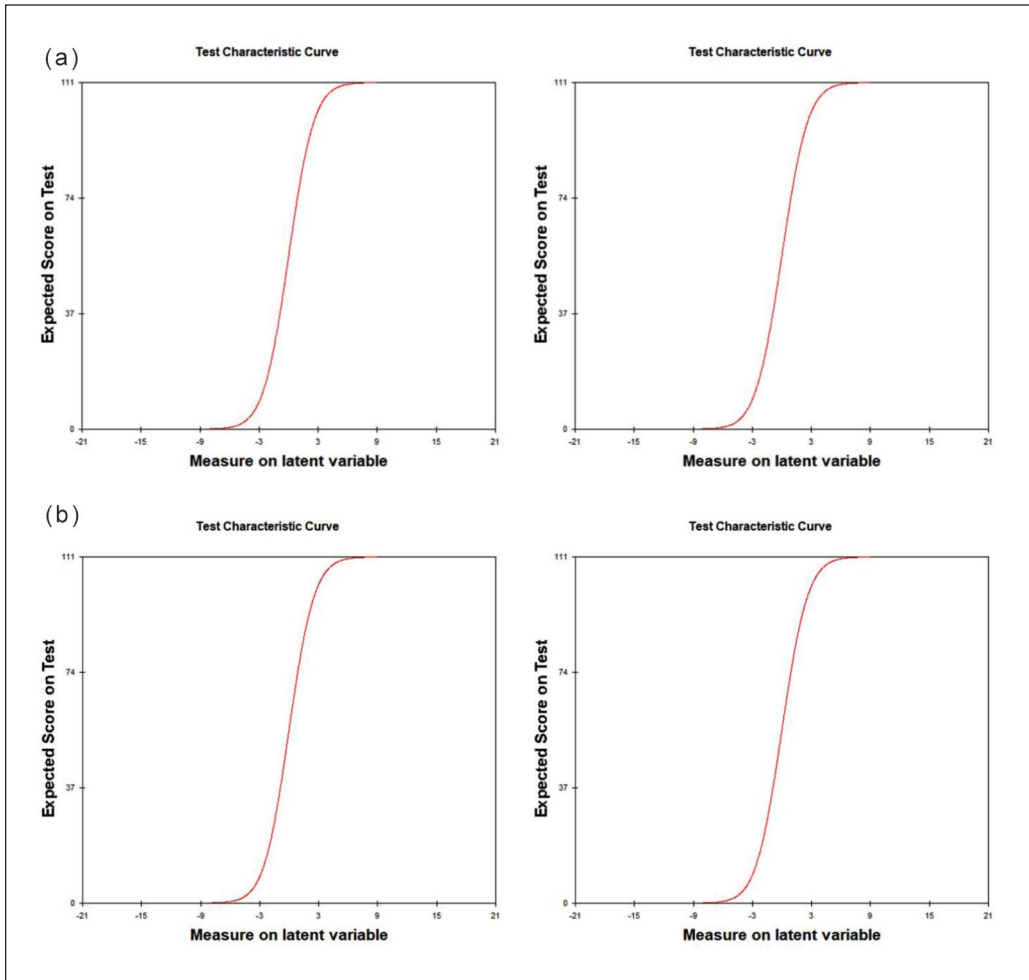*Note.* DIF = differential item functioning.

Figure 2. TCC for girls and boys, respectively: (a) Sample 1 and (b) Sample 2.
Note. TCC = test characteristic curve.

.005) in the second sample. The high correlation values implied negligible DIF results between gender groups on the test-level. As seen in Figure 2(a) and (b), TCCs are invariant across gender

groups for both samples, meaning that DIF did not impact overall test functioning practically across girls and boys.

For each ER item, between-group fit statistics and $t = $ ZSTD value were obtained to test the deviations from expected responses across ethnic groups. The hypothesis tested with between-group fit statistics was that an item showed no overall DIF across all ethnicity groups. In both samples, none of the between-group fit statistics value was significant. Significant between-group fit values would warrant further pairwise comparisons between pairs of ethnicity groups but that was not the case for each sample, suggesting that none of the ER items function differently for children from different ethnic groups.

Discussion

In this research, we aimed to cross-validate a web-based assessment designed to measure children's understanding of others' emotions. This study establishes evidence of construct validity for this ER assessment with two different samples. We evaluated several dimensions of construct validity, including assessment content, substantive validity, structural validity, responsiveness, and generalizability in two large samples. The results were consistent across the samples, supporting conclusion about the psychometric properties of the ER assessment. Consistent and plausible item fit results in both samples indicate that items fit Rasch model expectations. The analysis of dispersion of items along the latent trait continuum revealed good representation of item difficulties along a wide range of person abilities. Across samples, the empirical item difficulty hierarchies were consistent with each other. PCAs support a unidimensional structure and support the structural aspect of validity. Analyses of DIF and DTF across different gender and race groups provided evidence for generalizability across gender and ethnicity. A small number of items (six out of 111) displayed DIF across gender. The presence of these items nevertheless did not appear to have an untoward effect on overall test functioning, as correlations between Rasch scores with and without DIF items were approximately .99.

Significance

This study is the only one we are aware of to use a Rasch framework to validate an ER assessment and to evaluate the DIF of a facial ER task across gender and ethnicity. In two large samples, we found evidence supporting multiple forms of validity. Because item difficulties were designed to vary and targeted person abilities well, it is possible to integrate this bank of ER into an adaptive testing system that would efficiently yield reliable estimates of ER ability.

In addition, we found no evidence of meaningful DIF or DTF on the ER task. As a result, users of this ER assessment can be confident that item and total scores have the same relationship to ER skill, whatever their gender or ethnic background. Although this may be true of other facial ER assessments, no published work reports empirical evidence supporting that conclusion. As a result, it is not possible to know with confidence that other ER item and test scores reflect child competence the same regardless of group membership.

Limitations and Future Directions Implications

Despite evidence of construct validity, some limitations merit further examination. For example, it will be important to understand the source of DIF of the six items whose functioning differed for boys and girls. Because of the nature of the item content, item revision is not a viable remedy to eliminate DIF for these items. Nevertheless, understanding its source will help guide the development of DIF-free items.

In addition, person separation values were adequate, but could be improved by increasing the number of items targeting very high and very low ability respondents. Adding very easy and very hard items to the assessment will improve person separation index values and the usefulness of the assessment across the entire range of person abilities.

Finally, the ER assessment met Stevens's (2002) criteria for unidimensionality. However, items that loaded on the second and third contrasts were each from the same emotion. This suggests that although ER is mainly a singular skill, the ability to recognize different emotions may be at least partially separable skills. Future research should examine the extent to which the ability to infer different emotions is distinct using, for example, longitudinal designs examining the extent to which recognition of different emotions develops distinctly across childhood.

Use of the SELweb ER in Practice

SELweb's ER assessment has been developed and validated to address technical shortcomings and complement existing assessments in terms of variety and intensity of emotions assessed, sample characteristics, and generalizability of results. The instrument is designed to be administered in conjunction with assessments of related but partially distinct constructs, including social perspective-taking, social problem-solving, and self-control. Prior research has established that performance on SELweb, including its ER module, is positively associated with behavioral and academic functioning. The present study adds to that work by demonstrating that the items that are part of the ER assessment are largely free of bias. As a result, it is reasonable to interpret ER item and test score performance as having a comparable meaning for boys and girls and for children from different ethnic groups.

ORCID iD

Clark McKown [iD] https://orcid.org/0000-0001-9694-1179

References

Banerjee, R., & Watling, D. (2005). Children's understanding of faux pas: Associations with peer relations. *Hellenic Journal of Psychology*, *2*, 27-45.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*, 647-663.

Denham, S. A. (2006). Social-emotional competence as support for school readiness: What is it and how do we assess it? *Early Education and Development*, *17*, 57-33. doi:10.1207/s15566935eed1701_4

Dusenbury, L., Dermody, C., & Weissberg, R. P. (2018). *2018 state scorecard scan*. Collaborative For Academic, Social, and Emotional Learning. Retrieved from https://casel.org/wp-content/uploads/2018/02/2018-State-Scan-FINAL.pdf

Ekman, P., & Friesen, W. V. (1976). *Pictures of Facial Affect (POFA)*. Retrieved from https://www.paulekman.com/product/pictures-of-facial-affect-pofa/

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial action coding system*. Salt Lake City, UT: Research Nexus, Network Research Information.

Gratz, K. L., & Roemer, L. (2004). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the difficulties in emotion regulation scale. *Journal of Psychopathology and Behavioral Assessment*, *26*, 41-54.

Herba, C. M., Landau, S., Russell, T., Ecker, C., & Phillips, M. L. (2006). The development of emotion-processing in children: Effects of age, emotion, and intensity. *Journal of Child Psychology and Psychiatry*, *47*, 1098-1106.

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY* (2nd ed.). San Antonio, TX: Harcourt Assessment.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2003). *A user's guide to Winsteps/Ministeps: Rasch-model computer programs*. Chicago, IL: Mesa Press.

Linacre, J. M. (2008). The expected value of a point-biserial (or similar) correlation. *Rasch Measurement Transactions*, *22*(1), 1114.

Linacre, J. M. (2019). Winsteps (Version 4.3.4) [Computer software]. Beaverton, OR: Winsteps. Available from http://www.winsteps.com

Lipton, M., & Nowincki, S. (2009). The Social Emotional Learning Framework (SELF): A guide for understanding brain-based social emotional learning impairments. *The Journal of Developmental Processes*, *4*, 99-115.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2004). *Mayer-Salovey-Caruso Emotional Intelligence Test: Youth Version (MSCEIT: YV) item booklet*. Toronto, Ontario, Canada: Multi-Health Systems.

McKown, C. M. (2017). Social-emotional assessment, performance, and standards. *The Future of Children*, *27*, 157-178.

McKown, C.M. (2019). Reliability, factor structure, and measurement invariance of a web-based assessment of children's social-emotional comprehension. *Journal of Psychoeducational Assessment*, *37*, 435-449.

McKown, C. M., Russo-Ponsaran, N. M., Johnson, J. K., Russo, J., & Allen, A. (2016). Web-based assessment of children's social-emotional comprehension. *Journal of Psychoeducational Assessment*, *34*, 322-338.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *5*, 741-749.

Montirosso, R., Peverelli, M., Frigerio, E., Crespi, M., & Borgatti, R. (2010). The development of dynamic facial expression recognition at different intensities in 4- to 18-year-olds. *Social Development*, *19*, 71-92.

Murphy, K. R., & Davidshofer, C. O. (2004). *Psychological testing: Principles and applications*. Upper Saddle River, NJ: Pearson Prentice Hall.

Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior*, *18*, 9-35.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore, MD: Johns Hopkins University Press.

Singular Inversions. (2005). FaceGen main software development kit (Version 3.*1)*. Vancouver, British Columbia, Canada: Author.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*, 205-231.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, *51*, 541-565.

Smith, R. M. (2004). Detecting item bias with the Rasch model. *Journal of Applied Measurement*, *5*, 430-449.

Smith, R. M., & Plackner, C. (2009). The family approach to assessing fit in Rasch measurement. *Journal of Applied Measurement*, *10*, 424-437.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66-78.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Lawrence Erlbaum.

Tehrani-Doost, M., Noorazar, G., Shahrivar, Z., Banaraki, A. K., Beigi, P. F., & Noorian, N. (2017). Is emotion recognition related to core symptoms of childhood ADHD? *Journal of the Canadian Academy of Child and Adolescent Psychiatry/Journal de l'Academie Canadienne de Psychiatrie de l'enfant et de l'adolescent*, *26*, 31-38.

Thompson, R. A., & Goodman, M. (2009). Development of self, relationships, and socioemotional competence: Foundations for early school success. In O. A. Barbarin & B. Wasik (Eds.), *Handbook of child development and early education: Research to practice* (pp. 147-171). New York, NY: Guilford Press.

Tracy, J. L., & Robins, R. W. (2004). Putting the self into self-conscious emotions: A theoretical model. *Psychological Inquiry*, *15*, 103-125.

Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*, 221-261.

Weissberg, R. P., Goren, P., Domitrovic, C., & Dusenbury, L. (2012). *Effective social and emotional learning programs: Preschool and elementary school edition*. Chicago, IL: Collaborative For Academic, Social, and Emotional Learning.

Wolfe, E. W., & Smith, E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I—Instrument development tools. *Journal of Applied Measurement*, *8*, 97-123.

Wolfe, E. W., & Smith, E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement*, *8*, 204-234.

Wyssen, A., Lao, J., Rodger, H., Humbel, N., Lennertz, J., Schuck, K., . . . Munsch, S. (2019). Facial emotion recognition abilities in women experiencing eating disorders. *Psychosomatic Medicine*, *81*, 155-164.