Journal of Psychoeducational Assessment
2019, Vol. 37(4) 435–449

DOI: 10.1177/0734282917749682
journals.sagepub.com/home/jpa
SAGE

# Reliability, Factor Structure, and Measurement Invariance of a Web-Based Assessment of Children's Social-Emotional Comprehension

**Clark McKown**[1]

**Abstract**

The purpose of this study was to evaluate the psychometric properties and measurement invariance of a web-based, self-administered battery of assessments of social-emotional comprehension called "SELweb." Assessment modules measured children's ability to read facial expressions, infer others' perspectives, solve social problems, delay gratification, and tolerate frustration. In an ethnically and socioeconomically diverse sample of 4,419 children in kindergarten through third grade who completed SELweb: (a) scores from assessment modules exhibited moderate to high internal consistency and moderate 6-month temporal stability; (b) composite assessment scores exhibited high reliability; and (c) assessment module scores fit a theoretically coherent four-factor model that includes factors reflecting emotion recognition, social perspective-taking, social problem-solving, and self-control. In addition, the present study supports configural and metric invariance across time, sex, and ethnicity. Analyses suggest partial scalar invariance across time, sex, and, to a lesser degree, ethnicity.

**Keywords**

elementary school, reliability, validity, social-emotional learning, measurement equivalence

Social-emotional comprehension includes the ability to encode, interpret, and reason about social-emotional information. The conceptualization of social-emotional comprehension in this article draws on several theoretical traditions and hypothesizes four interrelated dimensions (Lipton & Nowicki, 2009). First, drawing on research on nonverbal communication, emotion recognition is defined as the ability to understand others' emotions from nonverbal cues (Nowicki & Duke, 1994; Pons, Harris, & de Rosnay, 2004). Second, drawing on research on children's theory of mind understanding, social perspective-taking is defined as the ability to interpret others' mental states (Wellman & Liu, 2004). Third, drawing on research on social information processing, social problem-solving is defined as the ability to reason about social problems

[1]Rush University Medical Center, Chicago, IL, USA

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

(Crick & Dodge, 1994; Denham, 2006). Finally, self-control includes the effortful control of attention, emotions, and behavior to achieve a goal (Duckworth, 2011).

Social-emotional comprehension is consequential. A large body of research suggests that emotion recognition, social perspective-taking, social problem-solving, and self-control are each associated with outcomes as wide ranging as self-esteem, locus of control, peer acceptance, physical health, substance use, income, socioeconomic status, single parenthood, and criminality (Blair & Raver, 2015; Blair & Razza, 2007; Crick & Dodge, 1994; Denham, 2006; Denham et al., 2012; Dubow, Tisak, Causey, Hryshko, & Reid, 1991; Duckworth & Seligman, 2005; Iyer, Kochenderfer-Ladd, Eisenberg, & Thompson, 2010; Izard et al., 2001; Lecce, Caputi, & Hughes, 2011; Moffit et al., 2010; Nowicki & Duke, 1994). In prior work, for example, McKown, Russo-Ponsaran, Allen, Johnson, and Russo (2016) found that the better early elementary-aged children scored on a measure of social-emotional comprehension, the more teachers reported that children engaged in socially skilled behavior and the less they engaged in problem behavior. Furthermore, social-emotional comprehension was positively associated with peer acceptance and both teacher reported and directly assessed reading and math skills, controlling for IQ.

## The Need for Direct Assessments of Social-Emotional Comprehension

Social-emotional comprehension includes cognitive and affective skills that may not have straightforward behavioral correlates. As a result, third-party raters must make a high level of inference, potentially limiting the validity of this form of assessment for measuring social-emotional comprehension. Self-report may not be well suited to assessing social-emotional skills because self-reported skill is only modestly correlated with skill level measured more objectively (Shrauger & Osberg, 1981), and children may respond in ways that reflect social expectations more than actual skill level (Crowne & Marlowe, 1960).

Social-emotional comprehension lends itself to direct assessment, in which children demonstrate their skill in a particular domain by solving challenging domain-relevant tasks. Ideally, direct assessments will have adequate construct coverage, high ease of use, the ability to conduct group administration, and appropriateness for a wide range of children. Existing direct assessments vary in domain and age coverage and population for which they are appropriate. Most require expertise to administer, score, and interpret.

## SELweb to Assess Social-Emotional Comprehension

To address the need for direct assessments of children's social-emotional comprehension that can be administered at large scale to the general population, the author and his colleagues created SELweb, a web-based system for measuring social-emotional comprehension in kindergarten through third grade. SELweb includes five assessment modules, one each to measure emotion recognition, social perspective-taking, and social problem-solving. Two additional modules assess dimensions of self-control—delay of gratification and frustration tolerance. SELweb's assessment modules use direct assessment.

Prior research suggests that direct assessment is a promising approach to assessing social-emotional comprehension. In four separate studies examining SELweb and other assessments, McKown, Allen, Russo-Ponsaran, and Johnson (2013) and McKown and colleagues (2016) found that (a) composite social-emotional assessment scores exhibited latent factor internal consistency reliabilities (Nunnally & Bernstein, 1994) averaging greater than .80; (b) social-emotional assessment observed scores fit a four-factor structure reflecting emotion recognition, social perspective-taking, social problem-solving, and self-control; (c) social-emotional comprehension factor scores demonstrated convergent and discriminant validity; and (d) controlling for IQ and demographic characteristics, performance on SELweb was positively associated with peer

acceptance, teacher report of social skills, and multiple indicators of academic achievement, and negatively associated with teacher report of problem behaviors. Evidence of the reliability and validity of a Spanish language version SELweb is consistent with these findings (Russo, McKown, Russo-Ponsaran, & Allen, 2018).

## Measurement Equivalence of SELweb

Prior studies included diverse samples of typically developing and clinic-referred children. Similarity in findings across measures, samples, and analyses suggests that these assessments yield reliable scores that are valid for understanding how well-developed children's social-emotional comprehension is. Nevertheless, no prior research I am aware of has directly evaluated the measurement equivalence of direct assessments of social-emotional learning (SEL). Evaluating SELweb's measurement equivalence will provide important information on the extent to which SELweb scores can be interpreted in the same way for children from different groups.

Measurement equivalence can be tested within a confirmatory factor analysis (CFA) model (Dimitrov, 2010; Millsap, 2011), by comparing nested CFA models with varying degrees of equality constraints. The most basic question about measurement equivalence is whether the factor structure is the same across groups (configural invariance). Assuming configural invariance assumptions are met, a second important question is whether factor loadings are equivalent for different groups (metric invariance). Metric invariance means that a one-unit change in the latent construct is reflected by the same change in the observed variables for all groups. Assuming metric invariance requirements are met, a third important question is whether latent intercepts are equivalent for different groups (scalar invariance). Scalar invariance means that, at a given level of the latent variable, people from different groups achieve the same score on the observed variables. Assuming scalar invariance assumptions are met, another question is whether residual item and factor variances and covariances are equal across groups (strict invariance).

This study examines the measurement equivalence of SELweb for boys and girls, for children from different ethnic groups, and across two administrations. Testing measurement equivalence for sex and ethnicity is important to ensure that the scores achieved by children from different groups have the same meaning. Testing measurement equivalence across time is important for two reasons. First, if SELweb is administered to the same child more than once, it will be important to ensure that testing familiarity or fatigue does not change the meaning of the obtained scores. Second, establishing measurement equivalence across time helps ensure that observed changes in SELweb scores over time reflect changes in what is being measured.

The first study goal was to evaluate the reliability and validity of SELweb modules and factor scores. Based on prior work (McKown and colleagues, 2016), I hypothesized that module score reliabilities would be between .70 and .80, that factor score reliabilities would exceed .80, and that module scores would reflect a four-factor structure that includes factors reflecting emotion recognition, social perspective-taking, social problem-solving, and self-control. The second study goal was to evaluate measurement equivalence of SELweb's factor structure across time, sex, and ethnicity. To do so, I tested a series of nested CFA models with increasingly stringent equality constraints to assess invariance across time, sex, and ethnicity (Millsap, 2011). Specifically, I tested configural, metric, and scalar invariance across time, sex, and ethnicity.

## Method

### *Participants*

The sample included 4,419 children from 20 schools in three urban and six suburban school districts in five states who were tested during 2014-2015. Mean age of participants was 7.5 years

**Table 1.** Sample Characteristics.

| Characteristic | *n* | (%) |
|---|---|---|
| Sex—Male | 2,211 | (50.0) |
| Low income | 2,469 | (55.9) |
| Ethnicity | | |
| White | 1,972 | (44.8) |
| Black | 575 | (13.0) |
| Latino | 1,409 | (31.9) |
| Asian | 254 | (5.7) |
| Other | 209 | (4.7) |
| Grade | | |
| K | 754 | (17.1) |
| 1 | 1,257 | (28.4) |
| 2 | 1,360 | (30.8) |
| 3 | 1,048 | (23.7) |
| Total | 4,419 | |

*Note.* Low-income estimates were taken from public records about the proportion of children eligible for free and reduced-price lunches, or whose families received public aid.

(*SD* = 1.1). Sample characteristics are summarized in Table 1. Ethnicity labels that were common across all districts included synonyms for White, Black, Latino, and Asian. Members of other ethnic groups are categorized in this study as "Other."

## *Procedures*

In all participating districts, school staff administered SELweb to all students in kindergarten through third grade to learn about their students' SEL skills. Districts received data on student SELweb performance. The investigators received de-identified data to evaluate SELweb's technical properties. The university's institutional review board (IRB) granted a waiver of informed consent to use de-identified SELweb for research purposes.

School personnel administered SELweb in one or two sessions in a room with several Internet-connected computers. All sessions were group administrations. To complete SELweb, children are logged in by an administrator. SELweb assessment modules are illustrated and narrated with pictorial forced-choice responses that respondents select with a mouse. Children wear headphones and complete SELweb autonomously. Scoring is described in McKown and colleagues (2016).

Total testing time was approximately 45 min. Kindergarten and first-grade students generally completed SELweb in two separate sessions of approximately equal length, typically within 1 week. Students in second and third grade nearly always completed the assessment in a single session. In one district, SELweb was administered in fall (October and November) and spring (April and May), and data from that district were used to estimate temporal stability. Mean time between administrations was 176 days (*SD* = 47.7 days).

## *Measures*

SELweb modules, described below, were designed to assess emotion recognition, social perspective-taking, social problem-solving, and self-control. Summary descriptions of SELweb's modules and item scoring rules are summarized in Table 2. Total scores on each item were summed across items within module.

**Table 2.** Description of SELweb Modules, Questions, and Item Scoring.

| Module | Stimulus | Items *n* | Question and response options | Item score | Possible range |
|---|---|---|---|---|---|
| Emotion recognition | Respondents view individual child faces and indicate emotion expressed. | 40 | What is the child feeling? Happy, sad, angry, scared, just ok. | 2 = correctly recognizes emotion; 1 = mistakes emotion for neutral; 0 = selects incorrect emotion | 0-80 |
| Social perspective-taking | Respondents listen to illustrated, narrated vignettes and answer questions. | 12 | Questions about character intention Illustrated, narrated forced-choice, four possible responses. | 2 = correct mental state inference 1 = correct answer, no mental state inference 0 = incorrect answer | 0-24 |
| Social problem-solving | Respondents hear illustrated, narrated vignettes involving either ambiguous provocation or peer entry. | 6 | Attribution Did the person do it to be mean? Yes or no; if yes, a little or a lot? | 2 = "no" 1 = "yes" and "a little" 0 = "yes" and "a lot" | 0-12 |
| | | 6 | Goal Preference How do you want it to turn out? Narrated forced choice. | 2 = positive goal; 1 = retribution goal; 0 = revenge goal | 0-12 |
| | | 6 | Solution Preference What would you do? Illustrated, narrated forced choice. | 2 = competent assertive; 1 = self-advocacy and ignoring; 0 = aggressive | 0-12 |
| Self-control: Delay of gratification | Children send illustrated rocket ships to space. One is fast. One is slower. One is very slow. | 10 | Children are told to get as many points as possible in 10 trials. | 3 = slowest rocket; 2 = medium rocket; 3 = fast rocket | 0-30 |
| Self-control: Frustration tolerance | Children view pairs of shapes and indicate whether they match. Several items are programmed to get "stuck." | 23 | Children click on a ""if shapes are the same and an "X" if they are different. Children have 90 s to complete. | 1 = correct response; 0 = incorrect response | 0-23 |

*Emotion recognition.* Six photographs of child faces with neutral facial expressions, including three girls, three boys, and two ethnic minorities, one Black girl and one mixed-race boy, were used to create the emotion module. The photographs were digitized with FaceGen software (Singular Inversions, 2005). FaceGen was then used to digitally manipulate the face images to produce emotion displays of happy, sad, angry, and frightened. For each face and emotion, 10 faces were created ranging from low- to high-intensity affect displays. From this item pool, five different test forms were created, each with 40 items. 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. During SELweb administration, after each face was presented, children clicked to indicate whether the face reflected happy, sad, angry, scared, or just okay.

*Social perspective-taking.* My colleagues and I created 12 illustrated and narrated vignettes, six of which assessed false belief understanding and six of which assessed children's ability to distinguish between what a speaker appears to say and what they really mean, as when they are sarcastic, lying, or hiding their feelings. After each story, children were asked a question whose correct answer required an accurate inferences about the story character's mental state.

*Social problem-solving.* We created 10 illustrated and narrated vignettes, five involving ambiguous provocation and five involving peer entry. After each vignette, children selected (a) the extent to which they felt a story character was hostile (not at all, a little, or a lot); (b) how they wanted things to turn out, indexing social goals (prosocial or retribution); and (c) what they would do (with a choice of actions that reflected competence, asking for help, ignoring, or walking away). We created five test forms with six vignettes each. Each form included three ambiguous provocation vignettes and three peer entry vignettes. Each vignette was included on three forms.

*Self-control.* We developed a choice-delay task (Kuntsi, Stevenson, Oosterlaan, & Sonuga-Barke, 2001) and a frustration-tolerance task (Bitsakou, Antrop, Wiersema, & Sonuga-Barke, 2006) described in Table 2. For the choice-delay task, children earned points for clicking on one of three animated rockets that would then travel to a planet. The slower and therefore more tedious the rocket ship, the more points were awarded. Children first completed a training phase in which the narrator explained the point value of each rocket, and children selected each rocket to understand each rocket's speed and point value. Children completed 10 trials.

For the frustration-tolerance task, children were told to identify whether or not pairs of shapes presented one after the other were identical and to get as many correct as possible in 90 s. For several items, the computer was programmed to get "stuck," thereby inducing mild frustration. Clicking on the response buttons results in no changes to the screen. In early work, the number of items correct was most strongly associated with other indicators of self-control. As a result, I use this score as the indicator of frustration tolerance.

## Results

Descriptive statistics, correlational analyses, and reliabilities were calculated using SPSS version 19.0 (IBM, 2010). CFAs, including multigroup analyses used to evaluate measurement noninvariance, were conducted with Amos version 17.0 (Arbuckle, 2008).

### *Normality Assumption*

To test multivariate normality, skewness and kurtosis were computed for all 10 observed scores, as summarized in Table 3. For nine out of 11 scores, the absolute value of the skewness

**Table 3.** Correlations Between Variables in the Model and Descriptive Statistics.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Happy | — | | | | | | | | | | |
| 2. Sad | .34 | — | | | | | | | | | |
| 3. Angry | .20 | .28 | — | | | | | | | | |
| 4. Scared | .31 | .49 | .39 | — | | | | | | | |
| 5. R-A | .15 | .28 | .22 | .34 | — | | | | | | |
| 6. F-B | .11 | .22 | .16 | .26 | .61 | — | | | | | |
| 7. PA | .11 | .19 | .12 | .22 | .39 | .37 | — | | | | |
| 8. Goal | .16 | .23 | .14 | .25 | .28 | .24 | .47 | — | | | |
| 9. Solution | .12 | .22 | .14 | .22 | .37 | .32 | .49 | .55 | — | | |
| 10. Delay | .05 | .15 | .10 | .17 | .38 | .34 | .24 | .17 | .24 | — | |
| 11. Frust | .16 | .27 | .19 | .30 | .42 | .34 | .24 | .26 | .26 | .25 | — |
| *M* | 14.5 | 14.8 | 15.2 | 15.6 | 8.6 | 7.8 | 9.0 | 11.0 | 8.7 | 22.7 | 16.8 |
| *SD* | 3.0 | 3.8 | 3.5 | 4.1 | 2.5 | 3.2 | 3.0 | 1.9 | 2.9 | 4.6 | 2.8 |
| Skewness | −1.2 | −1.0 | −1.1 | −1.2 | −0.6 | −0.5 | −0.9 | −2.5 | −1.0 | −0.1 | −2.1 |
| Kurtosis | 2.9 | 1.1 | 1.4 | 1.1 | −0.2 | −0.8 | 0.0 | 7.2 | 0.4 | −0.7 | 4.9 |

*Note.* All correlations significant at $p < .001$. R-A = real–apparent emotions; F-B = false beliefs; PA = social problem-solving positive attributions; Goal = social problem-solving goal preference; Solution = social problem-solving solution preference; Delay = delay of gratification; Frust = frustration tolerance.

was less than 2. Kurtosis of nine out of 11 variables had an absolute value less than 3. To evaluate the impact of these deviations from normality, a Monte Carlo simulation with 200 bootstrap samples was run. For this analysis, Amos drew random, and therefore normally distributed, samples with the same means, variances, and covariances as the observed data. The distribution of parameter estimates from the simulated data was compared with parameter estimates from the observed data. There were no statistically significant differences between parameters estimated from bootstrap samples and those estimated from observed data. This suggests that deviations from normality in this sample did not have a meaningful impact on parameter estimates.

## *Missing Data*

Complete data were available for all SELweb data except for social perspective-taking. Twenty-nine of 4,419 children did not have social perspective-taking scores (0.66%). Simulation studies with small amounts of missing data (<2% of cases) have found that mean substitution is equivalent to more complex procedures (McCartney, Burchinal, & Bub, 2006). Therefore, mean substitution was used to impute perspective-taking values for those 29 cases.

## *Descriptive Statistics and Correlations*

Descriptive statistics and Pearson correlations between variables in the CFA models are summarized in Table 3.

## *Reliability*

*Internal consistency.* The internal consistencies of observed scores and factor scores are summarized in Table 4. Internal consistencies of factor scores were estimated using procedures described by Nunnally and Bernstein (1994).

**Table 4.** Score Reliabilities.

| SELweb dimension and score | $r_{yy}$ | $r_{12}$ |
|---|---|---|
| Emotion recognition | .86 | .55 |
| Social perspective-taking | .79 | .79 |
| Social problem-solving | .88 | .64 |
| Self-control | .86 | .66 |
| **SELweb observed scores** | **α** | **$r_{12}$** |
| Emotion recognition | | |
| Happy | .74 | .35 |
| Sad | .74 | .45 |
| Angry | .76 | .48 |
| Scared | .70 | .52 |
| Social perspective-taking | .79 | .79 |
| False belief | .70 | .70 |
| Reality appearance | .63 | .71 |
| Social problem-solving | | |
| Positive attribution | .73 | .60 |
| Positive social goal | .72 | .46 |
| Positive solution selection | .82 | .57 |
| Self-control | | |
| Delay of gratification | .74 | .59 |
| Frustration tolerance | .85 | .54 |

*Note.* $r_{yy}$ = internal consistency reliability; $r_{12}$ = temporal stability reliability.

*Six-month stability.* Six-month measurement stability estimates are presented in Table 4. Because children were randomly assigned to emotion recognition and social problem-solving test forms, for those assessment modules, temporal stability estimates reflected a mix of alternate forms and test–retest reliability.

## *Factor Structure*

Prior research has found a four-factor model fits the SELweb observed scores. Consistent with that research, in the present study, the fit of a four model to the observed scores, as depicted in Figure 1, was excellent, $\chi^2(38) = 364.3$, $p < .05$, comparative fit index (CFI) = .97, root mean square error of approximation (RMSEA) = .044, 90% confidence interval (CI) = [.040, .048]. Accordingly, measurement invariance analyses reported below were based on this model.

In this and all models tested below, the chi-square tests of model fit were statistically significant, indicating a lack of fit between the data and the model. However, the chi-square test of model fit is sensitive to sample size, even when the fit of the data to the model is excellent (Brannick, 1995; Ullman, 2006). As a result, model fit was evaluated with CFI and RMSEA, indicators of model fit that are less sensitive to sample size. The configural model was judged to be a good fit to the data when CFI ≥ .95 and RMSEA ≤ .06 (Dimitrov, 2010). Metric invariance models for each grouping were compared with the configural model, and scalar invariance models for each grouping were compared with their respective metric invariance model. Criteria for rejecting the null hypothesis of metric and scalar invariance included a decrease in model fit from the less restrictive model of ≥.01 in CFI or an increase of ≥.015 in RMSEA (Chen, 2007).

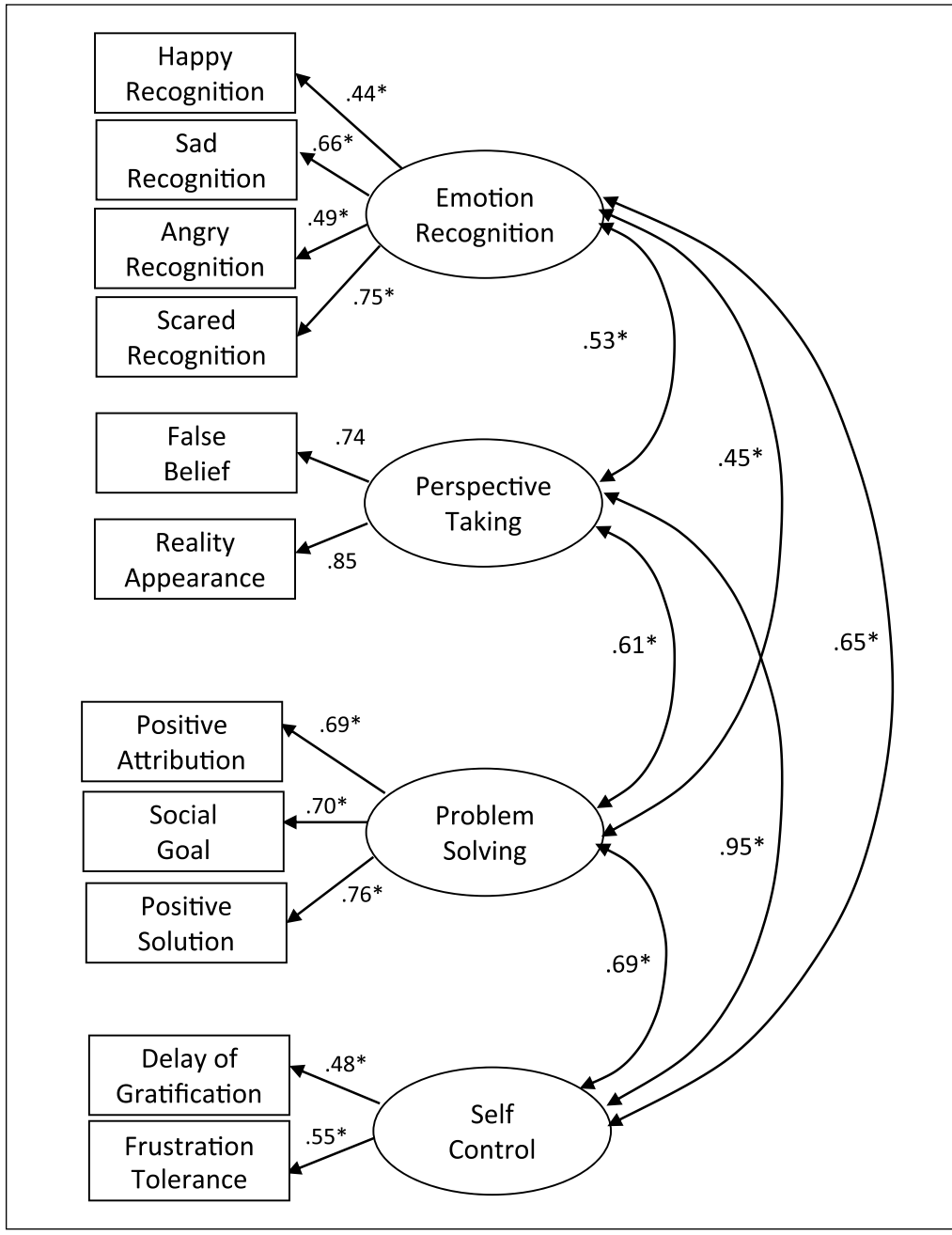


**Figure 1.** Four-factor model of social-emotional comprehension.
*Note.* Coefficients are standardized.
*$p < .05$.

## *Measurement Invariance*

Measurement invariance for each grouping variable (time, sex, and ethnicity) was tested by conducting multigroup CFA analyses and testing a series of nested models with increasingly

**Table 5.** Measurement Invariance Fit Statistics for Four-Factor Model.

| Model | Compared with | *df* | Δ*df* | $\chi^2$ | $\Delta\chi^2$ | RMSEA | ΔRMSEA | CFI | ΔCFI |
|---|---|---|---|---|---|---|---|---|---|
| Time invariance | | | | | | | | | |
| Configural | — | 76 | — | 368.8* | — | .031 | — | .975 | — |
| Metric | Configural | 83 | 7 | 409.4* | 40.6* | .031 | .000 | .972 | .003 |
| Scalar | Metric | 94 | 11 | 623.8* | 214.4* | .037 | .006 | .954 | .018 |
| Modified scalar[a] | Metric | 91 | 8 | 537.1* | 127.8* | .035 | .005 | .962 | .010 |
| Sex invariance | | | | | | | | | |
| Configural | — | 76 | — | 412.3* | — | .032 | — | .973 | — |
| Metric | Configural | 83 | 6 | 478.4* | 66.1* | .033 | .001 | .969 | .004 |
| Scalar | Metric | 94 | 11 | 738.2* | 259.8* | .039 | .006 | .949 | .020 |
| Modified scalar[b] | Metric | 92 | 9 | 611.5* | 133.1* | .036 | .003 | .959 | .010 |
| Ethnicity invariance | | | | | | | | | |
| Configural | — | 190 | — | 514.3* | — | .020 | — | .971 | — |
| Metric | Configural | 219 | 29 | 613.1* | 98.8* | .020 | .000 | .965 | .006 |
| Scalar | Metric | 262 | 43 | 1,428.2* | 815.1* | .032 | .012 | .896 | .069 |
| Modified scalar[c] | Metric | 243 | 24 | 746.7* | 143.6* | .022 | .001 | .955 | .010 |

*Note.* RMSEA = root mean square error of approximation; CFI = comparative fit index.
[a]Freed delay of gratification, false belief, and reality appearance.
[b]Freed delay of gratification and goal preference.
[c]Freed all false belief, White and Black reality appearance, Black, Hispanic, and other positive attribution; Black and Hispanic goal preference; Black and Hispanic solution preference; Black and Hispanic frustration tolerance; White, Black, and Asian delay of gratification.
*$p < .05$.

stringent equality constraints (Vandenberg & Lance, 2000) on the four-factor model. Ethnicity invariance analyses compared model fit between children identified as "White," "Black," "Latino," "Asian," or "Other."

*Configural invariance.* To test configural invariance, no equality constraints were imposed between groups on the four-factor model. A good fit with the unconstrained model supports configural invariance, or the equivalence of the factor structure across groups. As summarized in Table 5, across all configural invariance analyses, the models demonstrated an excellent fit to the data, with all CFI > .95 and all RMSEA < .05. In other words, for children who took SELweb twice, for boys and girls, and for children from different ethnic groups, the four-factor model fits all groups equally well.

*Metric invariance.* To test metric invariance, or the equivalence of the factor loadings across groups, equality constraints between groups were imposed on factor loadings. Imposing equality constraints on factor loadings across time, sex, and ethnicity led to a reduction in CFI < .01 and an increase in RMSEA < .002. This provided evidence of metric invariance across time, sex, and ethnicity. This means that across time, sex, and ethnicity, a one-unit change in latent variable is reflected by the same change in the observed scores.

*Scalar invariance.* To test scalar invariance, or the equivalence of factor intercepts across groups, additional equality constraints across groups were imposed on latent intercepts. With time and sex, the increase in RMSEA was <.015, consistent with a conclusion of scalar invariance, and the overall fit of the model remained excellent. However, across time and sex, the reduction in CFI for the scalar model CFI was >.01. The reductions in model CFI were .018 for time and .020 for sex.

Next, I freed the fewest equality constraints in the latent intercept to achieve a reduction in CFI < .01. To that end, I inspected modification indices to identify sources of scalar noninvariance. The key modification index was the expected drop in chi-square when freely estimating a given parameter. The parameter with the greatest modification index was freed first, and the model was rerun. This process was repeated until the discrepancy in CFI between the modified scalar model and the metric invariance model was reduced to <.01.

For time, when the delay of gratification, false belief, and reality-appearance intercepts were freed, the fit of the model was restored to equivalence with the metric invariance model, with a decline in model fit from the metric invariance model to the modified scalar model of .005 and .010 for RMSEA and CFI, respectively. For sex, when the delay of gratification task and goal preference intercept equality constraints were freely estimated, the fit of the model was restored to equivalence with the metric invariance model, with a decline in model fit from the metric invariance model to the modified scalar model of .003 and .010 for RMSEA and CFI, respectively. In the case of time and sex, SELweb therefore demonstrates partial scalar invariance.

The scalar invariance model for ethnicity resulted in a reduction in CFI of .069 to .90 and an increase in RMSEA of .012 to .032, meaning that for CFI but not RMSEA, imposing equality constraints across ethnicity on the latent intercepts resulted in a substantially poorer fit of the model to data than the metric invariance model. Sequentially freeing 14 of the 44 intercept equality constraints resulted in an improvement in model fit to CFI = .95 and RMSEA = .024, reducing the increase in RMSEA from the metric model to the modified scalar model to .004, and restoring the CFI to a value that reflects an excellent overall fit (Dimitrov, 2010). Thus, for ethnicity, a scalar invariance model in which 14 equality constraints were freely estimated partially met the criteria for scalar invariance (ΔRMSEA < .01). However, the CFI of this modified scalar invariance model was still .026 lower than the CFI from the metric model. Freeing an additional five equality constraints improved model fit to CFI = .955 and RMSEA = .022, reflecting a decline in model fit of .001 and .010 for RMSEA and CFI, respectively.

Because analyses found only partial support for the scalar invariance, more restrictive models, including strict invariance models, were not tested.

## Discussion

### *Summary and Interpretation*

The goals of this study were to evaluate key measurement properties of SELweb. A first goal of this study was to evaluate evidence of score reliability, based on evidence of internal consistency and temporal stability, and validity, based on evidence from the factor structure of the scores. Specific hypotheses related to this goal were based on prior field trials of SELweb and similar assessments (McKown et al., 2013, 2016).

In terms of reliability, consistent with prior findings, I hypothesized that the internal consistency reliability and temporal stability reliability of module scores would be moderate, and that factor score reliabilities derived from multiple correlated indicators would be substantially higher. Findings from the present study support these hypotheses. In general, score reliabilities—both internal consistency and temporal stability—at the level of the factor score were sufficiently high for the purposes of understanding student strengths and weaknesses in the areas assessed. In contrast, score reliability for the scores that were used as the indicator variables in the construction of those factor scores was low enough that they should be interpreted with caution when using SELweb to understand individual student strengths and needs.

In addition to internal consistency reliability, temporal stability coefficients were modest, raising important questions about SELweb's utility for detecting change over time. It is reassuring that across the 6-month test–retest interval, SELweb observed scores improved by

an average of .25 standard deviations, and the change in all observed scores was statistically significant. This suggests that SELweb scores are sensitive to the normative changes in social-emotional skill that unfolds over the course of a school year. Because temporal stability estimates were taken over a 6-month interval, rather than the traditional 2-week interval, test–retest reliability statistics computed for this study should be interpreted as a lower-limit estimate. In the future, it will be important to obtain test–retest reliability estimates over a shorter interval.

In terms of factor structure, consistent with prior findings, I hypothesized that the key indicator scores derived from SELweb would fit a four-factor model that includes latent variables reflecting emotion recognition, social perspective-taking, social problem-solving, and self-control. A confirmatory model supported this hypothesis. The overall fit of the model to the data was excellent, and factor loadings were consistently robust. In addition, covariances were generally moderate, suggesting that that the latent variables in the model are related but partially distinct.

One exception was the high covariance between the self-control and perspective-taking latent variables. This suggests that although social perspective-taking and self-control are conceptually distinct, they share a common underlying feature. Perhaps, for example, both reflect metacognitive skill. Nevertheless, because they are conceptually distinct constructs, I have modeled them as separate latent variables. Future research should investigate the nature of the common variance between these two seemingly different constructs that nevertheless covary highly.

A second goal of this study was to evaluate the measurement equivalence of SELweb across time, sex, and ethnicity. Specifically, this study evaluated the extent to which SELweb scores reflect the same underlying constructs on the same scale, with the same relationship between underlying skill level and observed score.

Analyses supported SELweb's configural invariance across time, sex, and ethnicity, meaning that observed scores reflect the same underlying factor structure across these ways of grouping respondents. Analyses also supported SELweb's metric invariance across time, sex, and ethnicity. This means that a one-unit change in the latent variable is associated with the same change on the observed scores across time, sex, and ethnicity. Furthermore, across time and sex, the data supported partial scalar invariance. Analyses provided less support for scalar invariance by ethnicity. This means that to a large degree for time and sex, and to a lesser degree for ethnicity, at a given level of skill on the latent variable, children from different groups achieve the same score on the observed variables.

These findings suggest that (a) for all groups, SELweb scores reflect the same underlying constructs (configural invariance); (b) for all groups, a change in skill level is reflected by the same change in observed scores; (c) for sex and repeated measurement, latent variable scores reflect similar observed scores; and (d) for ethnicity, latent variable scores may reflect different scores on the observed score. As a result, interpretation of mean differences between children from different ethnic groups on SELweb observed scores or composites should be made with caution.

An important area for future work on SELweb will be to identify and address the sources of ethnic noninvariance. Possibilities include ethnic differences in levels of engagement with the assessments, in effort applied to answering questions, and in interpretations of the meaning of assessment content. Until sources of noninvariance are identified and addressed, it will be important for users of SELweb to interpret mean differences between members of different ethnic groups with caution. Applying separate norms for each ethnic group may be an appropriate remedy for some applications.

### *Significance and Future Directions*

This extends prior theory and research on children's social-emotional comprehension. Much of the existing theory and empirical research focuses on a single social-emotional skill area, including emotion recognition (Nowicki & Duke, 1994), theory of mind (Wellman & Liu, 2004), social

problem-solving (Crick & Dodge, 1994), and self-control (Duckworth, 2011). The present study's findings are consistent with prior research in each of these areas suggesting that these skill areas are measurable dimensions of social-emotional comprehension that are correlated but partially distinct.

The present work builds on these largely separate lines of work by integrating important social-emotional skills into one conceptual framework, which was in turn used to design the SELweb modules. Findings from this study are consistent with multicomponent models of SEL that describe the processes by which children encode, interpret, and reason about social and emotional information (Collaborative for Academic, Social, and Emotional Learning, 2017; Crick & Dodge, 1994; Lipton & Nowicki, 2009; Salovey & Mayer, 1990). There is considerable common ground between the findings of this article and each of those models, and between the models themselves. Nevertheless, each emphasizes some social-emotional processes more than others. An important future direction for the field is, therefore, to clarify the commonalities and distinctions between models of SEL.

This work suggests important next steps in the practical application of SELweb and assessments like it. SELweb assesses social-emotional skills that are commonly taught in evidence-based SEL programs (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011; Jones et al., 2017). Ongoing field trials of SEL interventions are using SELweb as an outcome measure, and if SELweb is sensitive to intervention effects, it may be a useful program evaluation tool. Future work should also examine SELweb's usefulness as a formative tool by which educators can understand their students' strengths and needs and use that information to guide instruction and investment in programs.

The ultimate goal of SELweb, and assessments like it, is to inform instruction and intervention planning. In fact, SELweb assesses dimensions of SEL that are commonly addressed in evidence-based SEL curricula and clinical interventions. Ideally, then, teachers and other professionals will be able to use SELweb to guide instruction or intervention planning. The findings of this research and other studies of SELweb's psychometric properties suggest that it has many of the technical properties of just such an assessment. Further work to increase score reliability, eliminate sources of scalar noninvariance, and evaluate sensitivity to treatment effects will provide important additional information about SELweb's practical usefulness.

## Author's Note

The opinions expressed are those of the author and do not represent views of the Institute of Education Sciences or the U.S. Department of Education.

## Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Clark McKown has financial interests in xSEL Labs, Inc. which could potentially benefit from the outcomes of this research.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences through Grant R305A110143 to Rush University Medical Center.

## References

Arbuckle, J. L. (2008). *Amos* (Version 17.0). Chicago, IL: IBM.

Bitsakou, P., Antrop, I., Wiersema, J. R., & Sonuga-Barke, E. J. (2006). Probing the limits of delay intolerance: Preliminary young adult data from the Delay Frustration Task (DeFT). *Journal of Neuroscience Methods*, *151*, 38-44. doi:10.1016/j.jneumeth.2005.06.031

Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, *66*, 711-731. doi:10.1146/annurev-psych-010814-015221

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. doi:10.1111/j.1467-8624.2007.01019.x

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, *16*, 201-213

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464-504.

Collaborative for Academic, Social, and Emotional Learning. (2017). *Core SEL competencies*. Retrieved from http://www.casel.org/core-competencies/

Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin*, *115*, 74-101. doi:10.1037/0033-2909.115.1.74

Crowe, L. M., Beauchamp, M. H., Catroppa, C., & Anderson, V. (2011). Social function assessment tools for children and adolescents: A systematic review from 1988 to 2010. *Clinical Psychology Review*, *31*, 767-785. doi:10.1016/j.cpr.2011.03.008

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349-354. doi:10.1037/h0047358

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

Denham, S. A., Bassett, H. H., Thayer, S. K., Mincic, M. S., Sirotkin, Y. S., & Zinnser, K. (2012). Observing preschoolers' social-emotional behavior: Structure, foundations, and prediction of early school success. *Journal of Genetic Psychology*, *173*, 246-278.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*, 121-149. doi:10.1177/0748175610373459

Dubow, E. F., Tisak, J., Causey, J., Hryshko, A., & Reid, G. (1991). A two-year longitudinal study of stressful life events, social support, and social problem-solving skills: Contributions to children's behavioral and academic adjustment. *Child Development*, *62*, 583-599.

Duckworth, A. L. (2011). The significance of self-control. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 2639-2640. doi:10.1073/pnas.1019725108

Duckworth, A. L., & Seligman, M. E. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, *16*, 939-944.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*, 405-432. doi:10.1111/j.1467-8624.2010.01564.x

IBM. (2010). *IBM SPSS statistics for windows* (Version 19.0). Armonk, NY: Author.

Iyer, R. V., Kochenderfer-Ladd, B., Eisenberg, N., & Thompson, M. (2010). Peer victimization and effortful control: Relations to school engagement and academic achievement. *Merrill-Palmer Quarterly*, *56*, 361-387.

Izard, C., Fine, S., Schultz, D., Mostow, A., Ackerman, B., & Youngstrom, E. (2001). Emotion knowledge as a predictor of social behavior and academic competence in children at risk. *Psychological Science*, *12*, 18-23.

Jones, S., Brush, K., Bailey, R., Brion-Meisels, G., McIntyre, J., Kahn, J, . . . Stickle, L. (2017). *Navigating SEL from the inside out. Looking inside & across 25 leading SEL programs: A practical resource for schools and OST providers*. Retrieved from http://www.wallacefoundation.org/knowledge-center/Documents/Navigating-Social-and-Emotional-Learning-from-the-Inside-Out.pdf

Kuntsi, J., Stevenson, J., Oosterlaan, J., & Sonuga-Barke, E. J. (2001). Test-retest reliability of a new delay aversion task and executive function measures. *British Journal of Developmental Psychology*, *19*, 339-348. doi:10.1348/026151001166137

Lecce, S., Caputi, M., & Hughes, C. (2011). Does sensitivity to criticism mediate the relationship between theory of mind and academic achievement? *Journal of Experimental Child Psychology*, *110*, 313-331.

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

McCartney, K., Burchinal, M. R., & Bub, K. L. (2006). Best practices in quantitative methods for developmentalists. *Monographs of the Society for Research in Child Development*, *71*, 42-64.

McKown, C., Allen, A. A., Russo-Ponsaran, N. M., Johnson, J. K. (2013). Direct Assessment of Children's Social-Emotional Comprehension. *Psychological Assessment*, *25*, 1154-1166.

McKown, C., Russo-Ponsaran, N. M., Allen, A. A., Johnson, J., & Russo, J. (2016). Web-based direct assessment of children's social-emotional comprehension. *Journal of Psychoeducational Assessment*, *34*, 322-338. doi:10.1177/0734282915604564

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

Nowicki, S. Jr., & 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* (3rd ed.). New York, NY: McGraw-Hill.

Pons, F., Harris, P. L., & de Rosnay, M. (2004). Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European Journal of Developmental Psychology*, *1*, 127-152.

Russo, J., McKown, C., Russo-Ponsaran, N. M., Allen, A., (2018). Reliability and validity of a Spanish language assessment of children's social-emotional learning skills. *Psychological Assessment*, *30*, 416-421.

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition, and Personality*, *9*, 185-211.

Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin*, *90*, 322-351. doi:10.1037/0033-2909.90.2.322

Singular Inversions. (2005). *FaceGen main software development kit* (Version 3.1). Vancouver, British Columbia, Canada: Author.

Ullman, J. B. (2006). Structural equation modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment*, *87*, 35-50.

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

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*, 523-541.