Personality and Individual Differences xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Measuring social and emotional skills in elementary students: Development of self-report Likert, situational judgment test, and forced choice items

Dana Murano[a,*], Anastasiya A. Lipnevich[b], Kate E. Walton[a], Jeremy Burrus[a], Jason D. Way[a], Cristina Anguiano-Carrasco[a]

[a] ACT, Inc. Center for Social, Emotional, and Academic Learning, ACT, Inc., 500 ACT Drive, Iowa City, IA, USA
[b] Queens College and The Graduate Center, City University of New York. 65-30 Kissena Blvd, Flushing, NY, USA

ARTICLE INFO

Keywords:
Social and emotional learning (SEL)
Assessment
Big Five
Elementary
Situational judgment test
Forced choice

ABSTRACT

As social and emotional learning (SEL) continues to gain popularity, the need for high-quality social and emotional skill assessments also increases. We conducted two studies to develop and validate items to measure social and emotional skills in third, fourth, and fifth grade students. The Big Five personality framework served as an assessment framework for image-enhanced Likert items, situational judgment test items, and forced choice items. Results from Study 1 (n = 1047) provided promising reliability and validity evidence, as well as concrete recommendations for item revisions. Study 2 (n = 826) was conducted with a revised item pool and demonstrated improved reliability and validity. Taken together, results provided initial support that social and emotional skills can be validly and reliably measured in elementary-aged students using innovative item types.

1. Introduction

Social and emotional learning (SEL) continues to gain momentum in the 21st century educational sphere, and the need to reliably measure student social and emotional skills is steadily growing, especially for younger students who have yet to develop strong reading skills. We discuss challenges associated with measuring social and emotional skills in elementary-aged students and common approaches used in the field and describe the development of a comprehensive and innovative assessment of elementary student social and emotional skills.

1.1. Assessment of social and emotional skills

Social and emotional skills are defined as "individual characteristics that originate from biological predispositions and environmental factors, manifested as consistent patterns of thoughts, feelings, and behaviors, developed through formal and informal learning experiences, and that influence different outcomes throughout the individual's life" (John & DeFruyt, 2015, p. 4). SEL has been associated with improved student outcomes such as positive attitudes toward school, decreases in deviant behavior, and enhanced academic performance (Casillas et al., 2012; Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011).

Having high quality assessments of social and emotional skills is critical for several reasons. First, reliable and valid assessments are key in being able to measure and monitor student skill development and evaluate SEL interventions. Second, they can be used formatively by teachers to guide classroom practices and interventions, and by students to monitor their own skill development (Marzano, 2015; Murano, Martin, Burrus, & Roberts, 2018). Additionally, they can be used to identify at-risk students who could benefit from early intervention (Denham, 2015).

1.1.1. Common approaches to measuring social and emotional skills and their disadvantages

Social and emotional skills assessments vary greatly depending on theoretical frameworks that underlie them. The lack of shared definitions and conceptual frameworks presents a first challenge and has implications for measurement (Abrahams et al., 2019). Assessment frameworks aside, various methods such as student self-report and other-informant (i.e., parents, teachers) are frequently used to measure social and emotional skills in elementary-aged students, each with their strengths and shortcomings (Denham, 2015; Kankaraš, Feron, & Renbarger, 2019).

One shortcoming across these assessments is the reliance on Likert items (e.g., Abrahams et al., 2019). Typically, Likert items present students with statements to which they respond to by circling a point on a scale (e.g., disagree/agree; not like me/like me). These items, while convenient, efficient, and appropriate for low-stakes testing, have a

* Corresponding author.
E-mail address: dana.murano@act.org (D. Murano).

https://doi.org/10.1016/j.paid.2020.110012
Received 15 January 2020; Received in revised form 22 March 2020; Accepted 23 March 2020

variety of response biases that can impact the validity of scores obtained. First, they are easy to fake, which can result in inflated mean scores on items perceived as socially desirable (Viswesvaran & Ones, 1999; Ziegler, MacCann, & Roberts, 2012). They are also susceptible to reference bias, which describes a response pattern in which people from different regions, backgrounds, levels of education, or norm groups may answer a question differently because each person's reference standard is dependent on his or own unique life experience (Kankaraš, 2017). Several response pattern biases also exist. Extreme response style is the respondent's tendency to choose extreme response categories on an item. In contract, the midpoint response style bias is the tendency to systematically select response options toward the middle of the response scale. Acquiescence bias is the tendency to consistently agree with statements, regardless of item content. Each of these biases can affect scale reliability and validity (Kankaraš, 2017).

1.2. Innovative approaches to measuring social and emotional skills

There are alternative approaches that can mitigate shortcomings associated with Likert items (Abrahams et al., 2019; Lipnevich, MacCann, & Roberts, 2013) in the social and emotional domain. This paper focuses on two innovative item types: situational judgment test (SJT) and forced choice (FC) items.

1.2.1. Situational judgment tests

SJT items can be advantageous over Likert items and are recommended as an alternative assessment solution in this domain (Abrahams et al., 2019). There are many different formats for SJT items (e.g., pick one response, pick most/least likely response). In all SJTs, respondents are presented with a scenario that they could possibly encounter in their day-to-day lives. Most formats also present several plausible behavioral responses, and respondents respond to each option. Scenarios presented to respondents should be both age- and context-relevant. SJTs are advantageous over Likert items because they can more precisely measure nuanced constructs (Lipnevich et al., 2013). They are also more difficult to fake than traditional Likert items because the most socially desirable response option is not always clear (Hooper, Cullen, & Sackett, 2006). Additionally, they show high predictive and face validity in educational settings (Lievens & Sackett, 2012; Wang, MacCann, Zhuang, Liu, & Roberts, 2009). Despite these advantages, they also have several shortcomings; they are more cognitively taxing, often demonstrate multidimensionality, and have notably lower reliabilities (e.g., Kasten & Freund, 2016; Lipnevich et al., 2013).

1.2.2. Forced choice

Instead of the traditional Likert approach, in which one stimulus is presented at a time, FC items present two or more adjectives or statements to respondents. Respondents are asked to choose what describes them most (and least), which can be done by ranking, or by selecting a single option for most and a single option for least (Lipnevich et al., 2013). FC items can mitigate several biases associated with Likert items. They are more difficult to fake because participants cannot rate themselves highly on more than one statement; instead they must choose among them, which decreases the effect of any impression management on the part of the respondent (Salgado & Tauriz, 2014; Stark, Chernyshenko, & Drasgow, 2005). Additionally, FC items mitigate the issue of reference bias because, when responding to FC items, the respondent only makes comparisons at the trait level within themselves; there are no comparisons with others or other entities that are dependent on the individual's points of reference (Jackson, Wroblewski, & Ashton, 2000; Stark et al., 2005). Despite these advantages, they can also be more cognitively demanding, sometimes exhibit low validity estimates, and are difficult to score (Brown & Maydeu-Olivares, 2013; Dueber, Love, Toland, & Turner, 2019).

1.3. Assessment framework

The Big Five framework was selected as the assessment framework. Borrowed from personality psychology, this framework was neither decided upon nor created, but rather discovered via a lexical analysis of words in the English language, which resulted in five factors: extraversion, agreeableness, conscientiousness, neuroticism (emotional stability), and openness to experience (Allport & Odbert, 1936). The use of the Big Five framework has several advantages. First, its structure is empirically based, rather than developed by expert consensus or theory alone. Second, it is generalizable across cultures with the factor structure of the Big Five having been confirmed in replication studies across a variety of languages and cultures (e.g., McCrae & Terracciano, 2005; Schmitt, Allik, McCrae, & Benet-Martinez, 2007). Third, the five factors can serve as a framework to organize different social and emotional skills, which is particularly useful when considering the jingle-jangle fallacies commonly found within this field (Burrus & Brenneman, 2016; Roberts, Martin, & Olaru, 2015). For example, a recent paper analyzed 50 SEL frameworks and found that they included 748 social and emotional competencies (Berg, Nolan, Yoder, Osher, & Mart, 2019). Fourth, meta-analytic evidence supports the validity of the Big Five for predicting educational (Poropat, 2009) and workforce (Barrick, Mount, & Judge, 2001) outcomes. For these reasons, the Big Five has been recommended as a universal framework for social and emotional skills by many (e.g., Abrahams et al., 2019; John & DeFruyt, 2015; Kautz, Heckman, Diris, ter Weel, & Borghans, 2014; Kyllonen, Lipnevich, Burrus, & Roberts, 2014; Primi, John, Santos, & De Fruyt, 2016) and is currently being used as the organizing framework for the Organization for Economic Cooperation and Development (OECD) worldwide study on student social and emotional skills (Chernyshenko, Kankaraš, & Drasgow, 2018).

1.3.1. Big Five structure in elementary-aged students

Use of the Big Five as an assessment framework is supported by evidence showing that the five-factor structure emerges in elementary school-aged children. Several studies have confirmed that the broad factor-level structure of personality in childhood is very similar to the structure in adulthood. The five-factor structure consistently replicates, and the factors demonstrate internal consistency and validity across countries and age groups (Chernyshenko et al., 2018). Of particular interest for the current study, the five-factor structure replicated in the 8–11 age group across three studies with good model fit, with factor loadings on the target factor ranging from 0.51–0.91, and scale reliabilities ranging from $\alpha = 0.72$ to $\alpha = 0.95$ (Halverson et al., 2003; Mervielde & De Fruyt, 1999; Tackett et al., 2012). In one study, the neuroticism factor proved to be the most difficult to replicate in younger samples, with its items correlating highly with (dis)agreeableness items (Tackett et al., 2012). In other words, agreeableness and emotional stability factors may be less differentiated in children compared to adults.

1.3.2. Predictive validity of the Big Five in academic contexts

Social and emotional skills show strong predictive validity for a range of desirable outcomes in educational contexts. In his meta-analysis, Poropat (2009) reported correlations between GPA and Big Five factors at the primary education level. All factors significantly correlated with GPA as follows: conscientiousness ($r = 0.28$), agreeableness ($r = 0.30$), emotional stability ($r = 0.20$), openness ($r = 0.24$), and extraversion ($r = 0.18$). Predictive relationships with other variables are also fairly consistent between elementary school-aged students and older children and adults, with conscientiousness and openness to experience emerging as the best predictors of educational performance (Poropat, 2009).

Other key factors for school success are student mental health and general well-being. Emotional stability, agreeableness, and conscientiousness all have strong relationships with overall well-being,

with emotional stability emerging as the best predictor for overall mental health. Further, average correlations with life satisfaction range from 0.17 (openness to experience) to 0.30 (emotional stability), with all other values exceeding 0.20 (Chernyshenko et al., 2018). Additionally, personality is related to student perceptions and feelings about school environments. Positive attitudes toward school have been associated with emotional stability, agreeableness, and conscientiousness (Heaven, Mak, Barry, & Ciarrochi, 2002). Finally, school climate has been cited as being reciprocally related to student social and emotional skills. Positive climate enables students to develop their social and emotional skills, and socially and emotionally competent students contribute to a school's positive climate (Osher & Berg, 2017).

1.4. Present study

Given the current lack of assessments that draw upon a common and robust framework, the goal of the current study is to develop and validate an assessment using Likert, SJT, and FC items to measure elementary students' social and emotional skills. Using data from two iterations of the ACT® Tessera® elementary school pilot study, which uses the Big Five as an assessment framework and measures five broad social and emotional skills, we sought to determine if image-enhanced Likert, SJT, and FC items showed acceptable reliability and validity evidence. We piloted an initial item pool in Study 1, and data from Study 1 informed item revisions for new items piloted in Study 2. We focus on obtaining validity evidence based on internal structure and relations with other variables.

2. Study 1 method

2.1. Participants

Participants were 1047 students in third (n = 342), fourth (n = 411), and fifth (n = 293) grade from 12 elementary schools in geographically diverse locations throughout the United States. Consent was obtained, parents and students were given the opportunity to opt-out, and administrators arranged administration procedures for students in the target grade levels(s) within the school. The sample was 53.3% female and identified their ethnicity as: American Indian/Alaska Native (2.2%), Asian (1.2%), Black/African American (13.9%), Hispanic/Latino (3.2%), Native Hawaiian/Other Pacific Islander (0.5%), White (59.1%), or as identifying with two or more races (8.2%). The remaining students (11.7%) chose not to respond.

2.2. Materials

Each participant completed the pilot version of ACT Tessera for elementary school students. The assessment measured five social and emotional skills: Grit, Teamwork, Resilience, Curiosity, and Leadership. Table 1 defines each social and emotional skill and shows how they align to the Big Five framework. Likert, FC, and SJT items were written to capture the skill definitions described in Table 1. All item writers were subject matter experts and were either PhD students or held PhDs in psychology. Likert and FC items made use of images, each of which described an adjective related to its respective social and emotional skill. Images were gender neutral and were intended to increase engagement and minimize cognitive load for the students, given their young age. A cognitive laboratory study was conducted prior to the pilot with elementary students, and item revisions were made based on these results. In order to assess the validity of the social and emotional skill scales, the assessment also included scales on life satisfaction, attitude toward school, school climate, and self-reported GPA.

2.2.1. Likert items

Six Likert items measured each skill, resulting in 30 total Likert items. Each item included an image with an accompanying descriptive adjective (see Fig. 1 for an example). Respondents rated how well each of the adjectives described them on a 4-point scale (*Not like me at all*, *Kind of like me*, *Mostly like me*, *A lot like me*). Two negatively keyed items per scale were reverse scored and then the scale score for each social and emotional skills was derived by taking the mean score of the six items per scale.

2.2.2. SJT items

Two SJT items were administered to measure each social and emotional skill (with the exception of the Grit scale, which included four items to reflect two separate facets of conscientiousness: responsibility and perseverance). Each item contained a stem, which presented a developmentally relevant scenario the student would be likely to experience, and then five response options, each of which offered a different behavioral response to the scenario. Respondents rated how likely they would be to engage in each of the behavioral responses on a 4-point scale (*Would not do for sure*, *Might not do*, *Might do*, *Would do for sure*). Each response option was scored as a separate indicator of the skill, resulting in ten items contributing to each SJT scale score (20 items for the Grit scale). The directionality of item scoring was determined empirically based on the direction of the correlation with the item's respective Likert scale in addition to review of item content. Items that negatively correlated with their respective Likert scale score and were designated as negative displays of the trait by expert review were reverse-scored. Each scale score was derived by taking the mean score of each individual item per scale. A sample SJT item appears in Fig. 2.

2.2.3. FC items

The FC section consisted of 30 total items (six per skill) that were arranged into 10 triads. Each triad contained three items, each of which measured a different skill. Within each triad, two of the items were positively keyed and one was negatively keyed, with the intention of being fit to a multi-dimensional IRT model for obtaining forced choice scores (Brown & Maydeu-Olivares, 2013). The items that were used in the triads were the same images with accompanying adjectives that appeared in the Likert scales. In response to each triad, respondents selected which of the three adjectives was most like them and which was least like them. All IRT models that were fit to the data failed to converge, so an ipsative approach was used to compute scores. A rank order was first generated from the participant responses (*Most like me* = 3, *Not selected* = 2, *Least like me* = 1). Scale scores were then generated by taking the mean of each of the rank order values derived from how the respondent ranked the six respective items per scale, with

Table 1
Social and emotional skill definitions and Big Five alignment.

Social and Emotional Skill	Big Five factor	Skill definition *The extent to which a student's actions demonstrate…*
Grit	Conscientiousness	Persistence, goal striving, reliability, dependability, and attention to detail at school
Teamwork	Agreeableness	Collaboration, empathy, helpfulness, trust, and trustworthiness
Resilience	Emotional Stability	Stress management, emotional regulation, a positive response to setbacks, and poise
Curiosity	Openness to Experience	Creativity, inquisitiveness, flexibility, open mindedness, and embracing diversity
Leadership	Extraversion	Assertiveness, influence, optimism, and enthusiasm

Fig. 1. Study 1 sample Likert Item.

negatively-keyed items reverse-scored. An example triad appears in Fig. 3.

2.2.4. Life satisfaction

This scale consisted of seven Likert items that measured respondents' self-reported satisfaction with life (α = 0.75). This scale was modified to be appropriate for elementary school-aged students from Huebner's (1991) Student's Life Satisfaction Scale, which has been shown to have acceptable psychometric qualities. Respondents rated how much they agreed with each of the statements on a 4-point Likert scale (*Disagree a lot*, *Disagree a little*, *Agree a little*, *Agree a lot*). Sample items included "*I have a good life*" and "*I wish some things in my life were different*" (reversed item).

2.2.5. Attitudes toward school

This scale consisted of four Likert items that measured students' attitudes toward school (α = 0.77). This scale was adapted from PISA items that measured students' attitudes toward mathematics by changing language referring to "math" to "school" (OECD, 2012). Respondents rated how much they agreed with each of the statements on a 4-point Likert scale (*Disagree a lot*, *Disagree a little*, *Agree a little*, *Agree a lot*). Sample items included "*I learn things in school that are important*" and "*School is a waste of my time*."

2.2.6. School climate

This scale consisted of six items (α = 0.84). Items were adapted from the California Healthy Kids Survey (CHKS; California Department of Education, 2008), Respondents rated how much they agreed with each of the statements on a 4-point Likert scale (*Disagree a lot*, *Disagree a little*, *Agree a little*, *Agree a lot*). Sample items included "*I feel like I am a part of my school*" and "*There are adults at my school who care about me*."

2.2.7. Self-reported academic performance

Students were asked to report their academic performance in response to the following item: "Please rate how well you think you are doing in each subject." Students provided ratings for math, science, reading, and their overall performance in school on a 4-point Likert scale (*Not very well*, *Okay*, *Pretty well*, *Very well*).

2.3. Procedure

All students completed the assessment during school hours in classroom settings. Students were given unlimited time to complete the assessment and were encouraged to ask proctors for help if they needed assistance answering any items. 1050 students completed an online version of the assessment via Qualtrics and 305 students completed an identical version of the assessment in paper-and-pencil format.

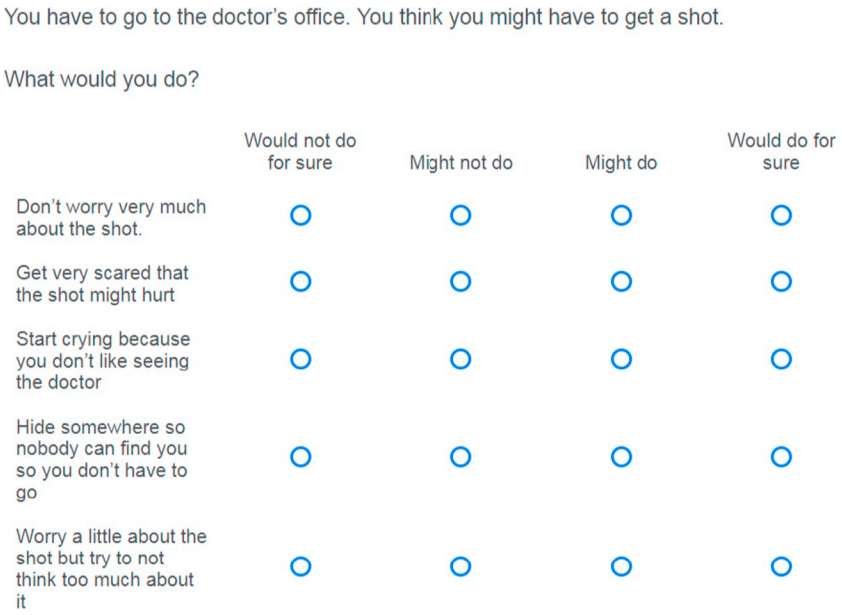

Fig. 2. Study 1 sample situational judgment test item.

Fig. 3. Study 1 sample forced choice item.

2.4. Analytic procedure

Prior to analyses, instances of low-quality responses were removed. Cases were excluded if they demonstrated any of the following response patterns: excessive missing data (> 20%), response time shorter than half the median testing time of the student's grade, variance < 0.1 on Likert or SJT items, or identical FC response patterns for all FC items. Of 1364 original cases in the data set, 1047 were used for analyses.

A confirmatory factor model was fit to determine if the Likert scale items fit the Big Five structure. Reliability analyses in line with classical test theory were also conducted for each Likert, SJT, and FC scale. Correlations were computed and examined to determine convergent, discriminant, and criterion-related validity. Last, a hierarchical regression was done in order to test the incremental validity of SJT and FC items over Likert items alone to improve the prediction of academic performance. All analyses were done using SPSS Version 23 or Mplus Version 7.

3. Study 1 results

3.1. Reliability and internal structure

3.1.1. Likert items

Scale scores and descriptive statistics were computed for all items. Cronbach's alpha was also computed for each Likert scale. Alpha values were as follows: Grit $\alpha = 0.66$, Teamwork $\alpha = 0.78$, Resilience $\alpha = 0.47$, Curiosity $\alpha = 0.72$, and Leadership $\alpha = 0.64$.

A confirmatory five-factor model was fit to the Likert data using weighted least squares estimation. The chi square test was significant ($\chi^2(395) = 3365.30$, $p < .001$) and fit statistics were as follows: RMSEA = 0.085 (CI: 0.082–0.087); CFI = 0.883=; .TLI = 0.871. Item-level factor loadings and descriptive statistics are reported in Table 2.

3.1.2. SJT items

Cronbach's alpha estimates were computed for each SJT scale using the ten behavioral responses per skill as individual items. Alpha values were as follows: Grit $\alpha = 0.80$, Teamwork $\alpha = 0.76$, Resilience $\alpha = 0.56$, Curiosity $\alpha = 0.42$, and Leadership $\alpha = 0.17$. Whereas some scales lacked internal consistency (i.e., Leadership), other scales such as Grit and Teamwork reached an acceptable level of reliability. These estimates are particularly notable considering that average internal consistency ratings for SJT scales average alphas of 0.57 (Campion, Ployhart, & MacKenzie, 2014). That is, SJT scales are expected to have lower internal consistency estimates than Likert items due to their complex nature.

A confirmatory five-factor model was fit to the SJT data to assess internal structure using weighted least squares estimation. Fit statistics were as follows: RMSEA = 0.083 (CI: 0.081–0.084); CFI = 0.718; TLI = 0.689.

3.1.3. FC items

Reliability estimates were computed for each scale. Cronbach's alpha estimates were as follows: Grit $\alpha = 0.49$, Teamwork $\alpha = 0.59$, Resilience $\alpha = 0.24$, Curiosity $\alpha = 0.48$, and Leadership $\alpha = 0.36$. These results should, however, be interpreted with caution given the distorted nature of reliability estimates resulting from ipsatively scored data (Meade, 2004).

Table 2
Study 1 means, standard deviations, and standardized factor loadings for Likert items.

Item	*M*	*SD*	Loading
Grit1	2.80	1.06	0.92
Grit2	2.77	1.00	0.92
Grit3(reversed)	3.37	0.88	0.85
Grit4	2.89	0.96	0.82
Grit5	3.22	0.83	0.76
Grit6(reversed)	3.21	0.83	0.59
Teamwork1	3.18	0.83	0.78
Teamwork2(reversed)	3.76	0.58	0.78
Teamwork3	3.53	0.69	0.76
Teamwork4	3.57	0.67	0.66
Teamwork5(reversed)	3.83	0.46	0.65
Teamwork6	3.52	0.67	0.31
Resilience1(reversed)	3.34	0.81	0.89
Resilience2(reversed)	3.15	0.84	0.88
Resilience3	2.71	0.93	0.85
Resilience4	2.79	1.02	0.59
Resilience5	2.70	0.98	0.28
Resilience6	2.83	1.00	0.12
Curiosity1	2.86	1.14	0.88
Curiosity2(reversed)	3.06	1.11	0.88
Curiosity3	2.91	1.12	0.69
Curiosity4	2.85	1.00	0.66
Curiosity5	3.21	0.96	0.61
Curiosity6(reversed)	3.53	0.88	0.55
Leadership1	3.09	0.85	0.86
Leadership2	2.88	1.08	0.84
Leadership3(reversed)	2.98	1.00	0.81
Leadership4	2.97	1.00	0.75
Leadership5(reversed)	3.15	1.04	0.73
Leadership6	3.36	0.84	0.54

Note. Estimates are standardized loadings on each respective.
Big Five factor when fit to a confirmatory five-factor model.

3.2. Relations to other variables

3.2.1. Convergent and discriminant validity

Correlations between all skills as measured by each item type were computed to examine convergent and discriminant validity. Table 3 contains correlations between all scales and across all item types. Convergent validity estimates averaged 0.35 for Grit, 0.39 for Teamwork, 0.25 for Resilience, 0.41 for Curiosity, and 0.16 for Leadership. Discriminant validity estimates averaged 0.25 for Grit, 0.25 for Teamwork, 0.24 for Resilience, 0.21 for Curiosity, and 0.07 for Leadership.

Table 4
Study 1 correlations between Likert, situational judgment test, and forced choice items with outcome measures.

	Academic performance	Life satisfaction	School attitude	School climate
Likert items				
Grit	0.21*	0.13*	0.23*	0.25*
Teamwork	0.30*	0.21*	0.32*	0.44*
Resilience	0.17*	0.17*	0.11*	0.17*
Curiosity	0.22*	0.14*	0.34*	0.37*
Leadership	0.09*	0.11*	0.04	0.03
Situational judgment test items				
Grit	0.25*	0.16*	0.34*	0.31*
Teamwork	0.21*	0.13*	0.38*	0.34*
Resilience	0.05	0.10*	0.14*	0.14*
Curiosity	0.07*	0.11*	0.21*	0.25*
Leadership	−0.03	0.01	−0.09*	−0.07*
Forced choice items				
Grit	0.13*	0.15*	0.21*	0.21*
Teamwork	0.15*	0.14*	0.24*	0.36*
Resilience	0.08*	0.12*	0.06	0.12*
Curiosity	0.16*	0.12*	0.31*	0.32*
Leadership	0.07*	0.10*	0.05	0.06*

* $p < .05$.

Recall, however, that the Resilience Likert scale and Leadership SJT scale had notably poor reliability estimates, and the Leadership scale was negatively correlated with all other scales.

3.2.2. Test-criterion validity

Table 4 shows the correlations between each scale and students' self-reported academic achievement, life satisfaction, attitude toward school, and perception of school climate. For relations between social and emotional skills and academic performance, correlations were expected to resemble those found by Poropat (2009) for students at the primary level. For Likert and SJT items, findings were in line with Poropat's findings. For FC items, Curiosity, Teamwork, and Grit had the strongest associations with academic performance, though with magnitudes were lower than those reported in Poropat ($r = 0.16$, $r = 0.15$, and $r = 0.13$, respectively). Resilience was expected to have the highest correlation with life satisfaction. This was not the case for any of the item types, but it should be noted that the Resilience Likert scale did now show acceptable reliability estimates, so these correlation coefficients cannot be

Table 3
Study 1 correlations among Likert, situational judgment test, and forced choice items.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Likert items														
1. Grit	–													
2. Teamwork	0.37*	–												
3. Resilience	0.32*	0.33*	–											
4. Curiosity	0.28*	0.42*	0.18*	–										
5. Leadership	0.02	0.14*	0.35*	0.07*	–									
Situational judgment test items														
6. Grit	**0.26***	0.35*	0.18*	0.31*	0.05	–								
7. Teamwork	0.14*	**0.32***	0.12*	0.26*	0.05	0.57*	–							
8. Resilience	0.08*	0.14*	**0.22***	0.12*	0.11*	0.30*	0.27*	–						
9. Curiosity	0.17*	0.30*	0.03	**0.35***	−0.01	0.25*	0.20*	0.13*	–					
10.Leadership	−0.04	−0.08*	−0.01	−0.06*	**−0.09***	−0.24*	−0.31*	−0.18*	−0.02	–				
Forced choice items														
11.Grit	**0.57***	0.17*	0.24*	0.21*	0.11*	**0.22***	0.13*	0.13*	0.09*	−0.05	–			
12.Teamwork	0.17*	**0.60***	0.21*	0.29*	0.11*	0.29*	**0.25***	0.10*	0.15*	−0.13*	0.16*	–		
13.Resilience	0.25*	0.27*	**0.42***	0.10*	0.14*	0.09*	0.08*	**0.10***	−0.01	−0.03	0.47*	0.37*	–	
14.Curiosity	0.16*	0.24*	−0.01	**0.66***	0.15*	0.29*	0.24*	0.12*	**0.22***	−0.07	0.26*	0.29*	0.05	–
15.Leadership	0.05	0.07*	0.12*	0.17*	**0.51***	0.08*	0.09*	0.08*	0.06*	**0.05**	0.27*	0.20*	0.15*	0.44*

Note. Bolded correlations indicate two scales are intended to measure the same skill and are therefore expected to be highest in magnitude.
* $p < .05$.

interpreted. Attitude toward school was expected to correlate most strongly with Resilience, Teamwork, and Grit. This was true, but Curiosity had higher correlation magnitudes than expected. Last, there was no expected pattern of correlations for school climate, other than that all skills should correlate positively with this variable. Correlations for each skill across scales averaged 0.26 for Grit, 0.38 for Teamwork, 0.14 for Resilience, 0.31 for Curiosity, and 0.01 for Leadership.

3.2.3. Incremental validity

A hierarchical linear regression was conducted to determine if the addition of SJT and FC items into a regression model accounted for variance in academic performance over and above Likert items alone. The five Likert scales were entered as the first step because this is the traditional method of measuring social and emotional skills, followed by the SJT scales in the second step, and FC scales as the last step. Likert items alone accounted for significant variance in academic performance ($R^2 = 0.11$, $F[5, 939] = 22.60$, $p < .01$). Adding SJT items accounted for additional variance ($R^2 = 0.13$, $\Delta R^2 = 0.02$, $\Delta F = 5.14$, $p < .05$). All three item types entered into the model accounted for 14% of the variance in academic performance ($\Delta R^2 = 0.01$, $\Delta F = 2.93$, $p < .05$). This provides evidence that including additional item types accounts for variance over and above that accounted for by Likert items alone.

4. Study 1 discussion

4.1. Summary of findings

In Study 1, we present reliability and validity evidence on Likert, SJT, and FC items designed to measure social and emotional skills aligned to the Big Five framework. This study provides moderate validity evidence of the new items yet highlights several places in which revisions can be made to improve the current item pool.

First, we consider reliability and validity evidence for image-enhanced Likert items. Likert scales aside from Resilience did reach or approach acceptable levels of reliability for low stakes, formative use cases (i.e., to provide feedback to students, to enable teachers to structure SEL instruction around students' scores). This shows promising evidence that image-enhanced items may be effective in engaging younger students in self-report items. However, the Likert items demonstrated poor model fit. Given that previous research supports the replicability of the Big Five structure with 8- to 11-year-old students (Halverson et al., 2003; Mervielde & De Fruyt, 1999; Tackett et al., 2012), this is likely attributable to the item content, rather than the theoretical model or factor structure. Some items did not load onto their intended factor, with several loadings well below 0.30. Item review identified several problematic items in terms of Big Five alignment (e.g., Likes school/does not like school, Brave, On-time, Angry), as well as the issue of negations used in items (e.g., Not shy, Not creative), which teachers consistently reported students struggling with. Better fit would likely result from revised item content.

Another potential issue is that Likert and FC items relied exclusively on adjectives matched with pictures. From a developmental perspective, the adjective format may have been too abstract, particularly for third grade students who are only eight years old (Piaget, 1964). Hence, being required to describe themselves using a single adjective may have been too developmentally complex. While adjectives were initially selected for their low reading load, it is possible to write short, contextualized sentences that describe discrete behaviors at an accessible reading level yet are not too abstract (e.g., "*I like to draw pictures*" compared to "*Artistic*"). Study 2 therefore makes use of sentences instead of adjectives.

In addition to these issues identified with Likert and FC items, we reviewed SJT content for similar issues with vocabulary and content alignment. We discovered content alignment issues with the Leadership SJT items, which made sense given the low scale reliability. We also hypothesized that language may have been too difficult for young students in some circumstances across scales and item types. We made revisions that addressed these issues and piloted the new item pool in Study 2.

4.2. Limitations and conclusions

In addition to the findings discussed above, there are several limitations to the current study. First, a small subset of the sample completed assessments using a paper-and-pencil format while the majority of participants took the survey online using Qualtrics. This source of method variance was not controlled for in the analyses. However, many of the cases obtained from paper-and-pencil format contained excessive amounts of missing data and were excluded from analyses. Second, negations for the scoring of SJT items were partially determined by the Likert items, and the Resilience scale did not approach an acceptable reliability coefficient. Reversals were reviewed based on content as well as the Likert scales, but is worth nothing that the SJT scoring could have been affected by unreliable Likert items.

Despite several concerns, the results presented did show promising evidence that image-enhanced Likert, SJT, and FC items can be administered to elementary-aged students to assess social and emotional skills. Moreover, the results showed moderate validity evidence and indicated that many items and scales did function as intended. This demonstrates potential for future self-report Big Five assessments to include multiple item types in order to obtain less biased measures of social and emotional skills in young students. However, revisions needed to be made to the item pool before use in operational settings, and additional validity evidence needed to be collected. Hence, items were revised based on the data collected in Study 1 and piloted in Study 2.

5. Study 2 method

5.1. Participants

The recruitment procedure was identical to that in Study 1. The same exclusion rules as in Study 1 were applied to the original sample of 925 students, resulting in 826 complete cases. Participants in the final sample included third ($n = 208$), fourth ($n = 323$), and fifth ($n = 295$) grade from seven elementary schools in the Midwest. In this sample, 52.7% of students reported being female, 44.8% reported being male, and 2.6% chose not to report their gender. Students in the sample identified their ethnicity as: American Indian/Alaska Native (6.5%), Asian (1.0%), Black/African American (1.5%), Hispanic/Latino (1.7%), Native Hawaiian/Other Pacific Islander (1.1%), White (57.3%), or as identifying with two or more races (6.4%). The remaining students (24.6%) chose not to respond.

5.2. Materials

Participants completed a revised item pool from Study 1. The same three item types were used with identical response scales, and the same five social and emotional skills were measured. The Flesch-Kinkaid reading level for the new form that used sentences instead of adjectives for Likert and FC items was 2.6.

5.2.1. Likert items

Six Likert items measured each skill, resulting in 30 total items. Many of the same images were used as in Study 1 but with several additional images. Additionally, images were described with full sentences instead of single adjectives (see Fig. 4). Items were also revised to remove any negations and words above a 3rd grade reading level.

5.2.2. SJT items

Problematic SJT items from Study 1 were reviewed and revised for inclusion in Study 2. In addition, data were analyzed further and it was determined that three behavioral response options could be used instead of five to reduce reading load and required testing time. Each of

Fig. 4. Study 2 sample Likert item.

the 10 revised SJT items contained a stem with three behavioral responses, each on a 4-point scale. Fig. 5 shows a sample SJT item.

5.2.3. FC items

The images from Study 1 were retained, but these items were also changed to sentences instead of adjectives to describe the images. Items were additionally reviewed and revised to remove negations, double negatives, and content deemed developmentally or contextually inappropriate. Fig. 6 shows an example.

5.2.4. Additional outcomes

In addition to the scales in Study 1, students completed the BFI-10, a brief measure of the Big Five factors (BFI-10; Rammstedt & John, 2007).

5.3. Procedure

All participants completed the assessment on Qualtrics. All other administration and analytic procedures followed those described in Study 1.

6. Study 2 results

6.1. Reliability and internal structure

6.1.1. Likert items

Cronbach's alpha was computed for each Likert scale, all of which reached an acceptable level of reliability: Grit $\alpha = 0.74$, Teamwork $\alpha = 0.80$, Resilience $\alpha = 0.76$, Curiosity $\alpha = 0.72$, and Leadership $\alpha = 0.73$.

The Likert data were fit to a five-factor confirmatory model using weighted least squares estimation. The chi square test was significant ($\chi^2(395) = 2379.04$, $p < .01$) and fit statistics indicated acceptable model fit (RMSEA = 0.078 [CI: 0.075–0.081]; CFI = 0.913; TLI = 0.904). Item-level factor loadings and descriptive statistics are reported in Table 5. Both fit statistics and alpha values showed improvements over those reported in Study 1.

6.1.2. SJT items

Cronbach's alpha estimates were also computed for each SJT scale treating the six items per skill as individual items. Alpha values were as follows: Grit $\alpha = 0.62$, Teamwork $\alpha = 0.60$, Resilience $\alpha = 0.51$, Curiosity $\alpha = 0.39$, and Leadership $\alpha = 0.55$. Overall, reliability estimates were lower than those in Study 1, but still demonstrated acceptable reliability evidence for SJT items. This was expected given the reduction in number of items, and desirable because of shortened test time.

Each response option was treated as an individual item and fit to a five-factor confirmatory model using weighted least squares estimation. The chi square test was significant ($\chi^2(395) = 2471.95, p < .001$) and fit statistics we as follows: RMSEA = 0.080 [CI: 0.077–0.083]; CFI = 0.809; .TLI = 0.801). Model fit improved with the reduced SJT scales.

6.1.3. FC items

Cronbach's alpha estimates for ipsatively scored forced choice scales were as follows: Grit $\alpha = 0.58$, Teamwork $\alpha = 0.58$, Resilience $\alpha = 0.39$, Curiosity $\alpha = 0.54$, and Leadership $\alpha = 0.49$. Alpha values improved compared to those in Study 1.

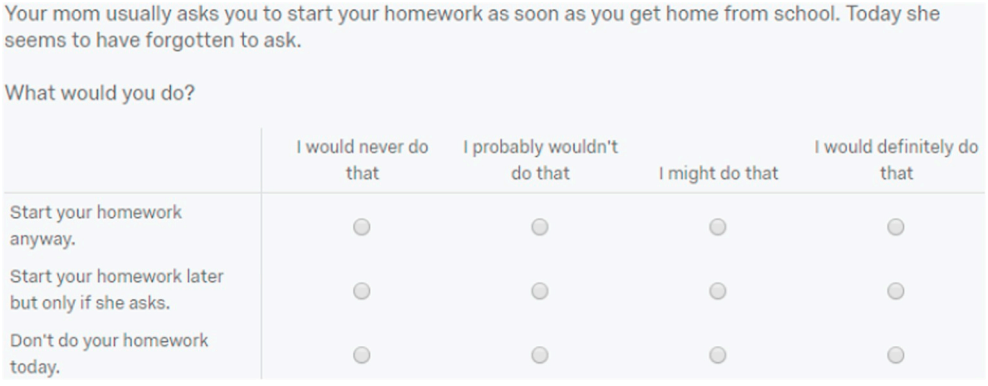

Fig. 5. Study 2 sample situational judgment test item.

Fig. 6. Study 2 sample forced choice item.

6.2. Evidence based on relations to other variables

6.2.1. Convergent and discriminant validity

Correlations between all scales across all item types were computed to examine convergent and discriminant validity and are reported in Table 6. Convergent validity estimates averaged 0.48 for Grit, 0.55 for Teamwork, 0.49 for Resilience, 0.48 for Curiosity, and 0.56 for Leadership. Discriminant validity estimates averaged 0.41 for Grit, 0.42 for Teamwork, 0.37 for Resilience, 0.38 for Curiosity, and 0.41 for Leadership. Though convergent validity improved, and convergent validity exceeds discriminant for each scale, inter-scale correlations increased from Study 1, resulting in weaker evidence of discriminant validity.

Convergent and discriminant validity were also evaluated using the BFI-10 scale (see Table 7). Across item type, convergent validity estimates averaged 0.40 for Grit, 0.41 for Teamwork, 0.40 for Resilience, 0.17 for Curiosity, and 0.26 for Leadership. Discriminant validity estimates averaged 0.29 for Grit, 0.29 for Teamwork, 0.27 for Resilience, 0.08 for Curiosity, and 0.12 for Leadership.

6.2.2. Test-criterion validity

Additionally, correlations were computed between scales for each item type and key outcome variables that were self-reported by students and reported in Table 7. Across the three item types, Grit was most highly correlated with academic performance (average $r = 0.36$), followed by Teamwork (average $r = 0.29$), with both magnitudes surpassing Poropat's and those found in Study 1. Similarly to in Study 1, Resilience still did not demonstrate the highest correlation with life satisfaction; Teamwork had the highest correlation magnitude ($r = 0.28$). Attitude toward school was expected to correlate most strongly with Resilience, Teamwork, and Grit. This was true across item types, and similarly to Study 1, Curiosity had higher correlations than expected ($r = 0.33$). Although Chernyshenko et al. (2018) did not report Curiosity as relating significantly with attitudes toward school, this makes sense theoretically. We know that openness correlates most highly with cognitive ability (e.g., Chernyshenko et al., 2018), and students who do well in school will likely also have positive attitudes toward school. All skills correlated positively with school climate across scales, with correlations averaging 0.34 for Grit, 0.45 for Teamwork, 0.31 for Resilience, 0.31 for Curiosity, and 0.35 for Leadership. All correlations were higher than those in Study 1.

6.2.3. Incremental validity

With Likert items alone, the five social and emotional skills accounted for significant variance in academic performance ($R^2 = 0.20$, $F[5, 730] = 36.94$, $p < .01$). Adding SJT items accounted for additional variance ($R^2 = 0.23$, $\Delta R^2 = 0.03$, $\Delta F = 5.67$, $p < .05$). When all three item types were entered into the model, they accounted for 26% of the variance in self-reported academic performance ($\Delta R^2 = 0.02$, $\Delta F = 4.64$, $p < .05$). The items in Study 2 accounted for 12% more variance in self-reported GPA than the items in Study 1.

7. Study 2 discussion

7.1. Summary of findings

Overall, the results from Study 2 show many improvements from those in Study 1. In terms of reliability, Likert and FC items improve greatly. SJT reliability estimates decreased slightly, but are still

Table 5
Study 2 means, standard deviations, and standardized factor loadings for Likert items.

Item	*M*	*SD*	Loading
Grit1	3.58	0.71	0.95
Grit2	3.55	0.71	0.88
Grit3	3.00	0.95	0.87
Grit4	3.20	0.90	0.85
Grit5	3.08	0.91	0.64
Grit6	2.84	0.95	0.57
Teamwork1	3.53	0.72	0.95
Teamwork2	3.52	0.70	0.94
Teamwork3	3.38	0.78	0.94
Teamwork4	3.35	0.78	0.93
Teamwork5	3.27	0.79	0.89
Teamwork6	2.83	1.02	0.79
Resilience1	2.71	0.99	0.91
Resilience2(reversed)	3.00	1.07	0.84
Resilience3	2.71	0.99	0.83
Resilience4(reversed)	2.82	0.97	0.83
Resilience5	2.65	0.92	0.82
Resilience6	2.39	1.03	0.40
Curiosity1(reversed)	3.55	0.81	0.94
Curiosity2	3.26	0.89	0.90
Curiosity3	3.04	0.95	0.69
Curiosity4	2.72	0.96	0.66
Curiosity5(reversed)	3.41	0.95	0.65
Curiosity6	3.43	0.86	0.60
Leadership1	3.15	0.90	0.89
Leadership2	2.72	1.01	0.88
Leadership3	3.23	0.86	0.86
Leadership4	2.99	0.97	0.82
Leadership5	2.79	1.11	0.71
Leadership6	2.81	1.07	0.60

Note. Estimates are standardized loadings on each respective.
Big Five factor when fit to a confirmatory five-factor model.

acceptable for this item type, and the shortened items are advantageous for this population due to reduced reading load and testing time. Model fit for Likert and SJT items also improved. Convergent, criterion, and incremental validity estimates increased across item types, though discriminant validity decreased.

7.2. Limitations and conclusions

Discriminant validity correlations improved when using the BFI-10 over the inter-scale matrix to evaluate discrimination. This was likely due to high inter-scale correlations between scales. With longer item sentence structure, student reading level could be a contributing factor to increased inter-scale correlations and less evidence of discriminant validity. Student reading level could be a confounding variable contributing to the scales being more highly correlated, and should be examined in future studies.

However, we also observed low correlations between all scales and the extraversion and openness scales of the BFI-10. While short to administer, the BFI-10 only contains two items per scale, and vocabulary may have additionally been too complex for students in this age group. More robust scales should be included for evaluating construct validity instead of the BFI-10, which includes only two items per factor. Including a longer Big Five measure validation for use with elementary-aged students would help gauge discrimination more accurately.

Overall, the Study 2 item pool shows great improvements from

Table 7
Study 1 correlations between Likert, situational judgment test, and forced choice items with outcome measures.

Scale	AP	LS	SA	SC	BFI_C	BFI_A	BFI_ES	BFI_O	BFI_E
Likert Items									
Grit	**0.44***	0.34*	0.40*	0.44*	**0.43***	0.38*	0.32*	0.10*	0.09*
Teamwork	0.32*	0.34*	0.37*	0.50*	0.35*	**0.43***	0.32*	0.07*	0.17*
Resilience	0.28*	0.31*	0.20*	0.33*	0.40*	0.31*	**0.54***	0.05	0.08*
Curiosity	0.26*	0.18*	0.46*	0.39*	0.29*	0.31*	0.17*	**0.16***	0.11*
Leadership	0.28*	0.28*	0.30*	0.40*	0.36*	0.31*	0.41*	0.11*	**0.27***
Situational judgment test items									
Grit	0.32*	0.19*	0.40*	0.30*	**0.35***	0.32*	0.19*	0.11*	0.11*
Teamwork	0.20*	0.22*	0.42*	0.42*	0.26*	**0.37***	0.19*	0.09*	0.06
Resilience	0.28*	0.23*	0.39*	0.36*	0.29*	0.32*	**0.21***	0.10*	0.16*
Curiosity	0.23*	0.15*	0.22*	0.24*	0.24*	0.23*	0.19*	**0.20***	0.21*
Leadership	0.22*	0.23*	0.28*	0.35*	0.23*	0.23*	0.26*	0.04	**0.20***
Forced choice items									
Grit	0.33*	0.29*	0.34*	0.32*	**0.39***	0.32*	0.32*	0.09*	0.10*
Teamwork	0.32*	0.29*	0.30*	0.40*	0.32*	**0.38***	0.31*	0.11*	0.16*
Resilience	0.30*	0.23*	0.17*	0.25*	0.31*	0.26*	**0.43***	0.06	0.10*
Curiosity	0.25*	0.15*	0.35*	0.33*	0.25*	0.29*	0.22*	**0.15***	0.14*
Leadership	0.24*	0.23*	0.29*	0.33*	0.31*	0.30*	0.32*	0.13*	**0.27***

Note. AP = Academic Performance, LS Life Satisfaction, SA = School Attitude, SC = School Climate, BFI_C = Conscientiousness, BFI_A = Agreeableness, BFI_ES = Emotional Stability, BFI_O = Openness, BFI_E = Extraversion.
Bolded correlations indicate two scales are intended to measure the same skill and are therefore expected to be highest in magnitude.
* $p < .05$.

Table 6
Study 2 correlations among Likert, situational judgment test, and forced choice items.

Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Likert items														
1. Grit														
2. Teamwork	0.59*													
3. Resilience	0.53*	0.54*												
4. Curiosity	0.48*	0.44*	0.34*											
5. Leadership	0.52*	0.65*	0.51*	0.47*										
Situational judgment test items														
6. Grit	**0.40***	0.28*	0.26*	0.35*	0.24*									
7. Teamwork	0.35*	**0.46***	0.28*	0.38*	0.37*	0.41*								
8. Resilience	0.42*	0.42*	**0.35***	0.41*	0.36*	0.38*	0.50*							
9. Curiosity	0.23*	0.26*	0.22*	**0.32***	0.29*	0.28*	0.26*	0.31*						
10. Leadership	0.32*	0.42*	0.28*	0.35*	**0.56***	0.23*	0.36*	0.34*	0.30*					
Forced choice items														
11. Grit	**0.56***	0.29*	0.42*	0.33*	0.31*	**0.33***	0.25*	0.32*	0.19*	0.26*				
12. Teamwork	0.38*	**0.63***	0.44*	0.29*	0.46*	0.26*	**0.40***	0.39*	0.22*	0.30*	0.29*			
13. Resilience	0.42*	0.41*	**0.62***	0.21*	0.36*	0.21*	0.21*	**0.34***	0.17*	0.18*	0.55*	0.50*		
14. Curiosity	0.31*	0.31*	0.22*	**0.63***	0.38*	0.31*	0.34*	0.40*	**0.34***	0.29*	0.32*	0.35*	0.21*	
15. Leadership	0.36*	0.42*	0.36*	0.43*	**0.56***	0.29*	0.34*	0.32*	0.30*	**0.37***	0.43*	0.44*	0.33*	0.59*

Note. Bolded correlations indicate two scales are intended to measure the same skill and are therefore expected to be highest in magnitude.
* $p < .05$.

Study 1 and supports the use of a multi-method approach to assessing social and emotional skills in elementary-aged students. However, FC and SJT scales can still be improved for reliability, and discriminant validity issues addressed through further revisions.

8. General discussion

8.1. Summary of findings

Taken together, these studies provide promising evidence for the development of Big Five-based items measuring social and emotional skills for third, fourth, and fifth grade students. The progression of this item pool represents item innovation in the social and emotional domain, as to our knowledge, no existing social and emotional skill measures make use of SJT or FC items with elementary students.

One robust conclusion is that reliability and validity improved from Study 1 to Study 2. One major modification was the use of sentences instead of adjectives in both Likert and FC items, which likely contributed to these improvements. Those developing items for students in this age range may be advised to develop items in the form of sentences rather than adjectives.

Another notable findings is that across both studies and the three item types, Teamwork emerged as: a) an internally consistent scale across item types, b) a strong predictor of all related outcomes, including GPA, and c) was highly correlated with all other scales and item types. This finding, paired with Poropat's (2009) meta-analytic finding that agreeableness has the highest correlation with GPA in primary education and lower magnitudes in secondary or tertiary levels of education, may suggest that agreeableness is a developmentally relevant skill at this age. Accordingly, Teamwork may be the most important factor for academic success at this age, as opposed to Grit, which is consistently linked with academic success after students enter secondary and post-secondary contexts.

8.2. Limitations

One limitation is that an ipsative approach was used to score FC items in both studies. This approach violates several assumptions of classical test theory, and as a result, reliability estimates are distorted. Though an alternative, IRT-based method to scoring FC items does exist and can be used in order to obtain normative score estimates (Brown & Maydeu-Olivares, 2013), all models fit to these data failed to converge. This was not surprising given recent issues with model convergence, particularly with models including < 30 traits (Burkner, Schulte, & Holling, 2019). Other problematic findings with this model include limited, and in some cases, decreased predictive, discriminant, and convergent validity with IRT scores over forced choice scores (Fisher, Robie, Christiansen, Speer, & Schneider, 2019; Walton, Cherkasova, & Roberts, 2019). Considering these issues and recent findings that ipsative and IRT-derived scores are often highly correlated (Walton et al., 2019), we felt that using an ipsative scoring approach was justified in order to evaluate FC scale validity. Future studies can include all positively keyed items, rather than one negatively keyed item per scale, in order to match items on social desirability.

Another limitation across both studies was the use of self-reported criterion variables. In particular, the only measure of student academic performance was reported by students. Criterion and incremental validity estimates may have varied given school-reported grades. Other informant ratings of student behavior and social and emotional skills could also be obtained to collect more robust validity evidence.

8.3. Future research

Future research could involve the development of a unified score containing scores from the three item types combined. Each item type has its own unique set of strengths and weaknesses. Use of multiple item types to create a combined score inclusive of multiple item types can mitigate biases of each item type, resulting in a score that is more valid and less biased than a score generated from a single method alone (Kankaraš et al., 2019). Moreover, each item type provides some unique contribution to the prediction of outcomes as shown by the incremental validity results. Currently, one such approach to measuring social and emotional skills exists (ACT, 2018), but is only available for middle and high school students. Results from ACT Tessera, which combined Likert, FC, and SJT items, show improved predictive validity of social and emotional skill scores in student GPA, increased reliability over Likert-based scores alone, and mitigation of faking and other response biases through the use of FC items (ACT, 2018; Anguiano-Carrasco, Walton, Murano, Burrus, & Way, 2018). Although computing a unified score was not appropriate in this study considering the unacceptable reliabilities and lack of convergence and discrimination of several scales across method types, future studies could aim to do so. Future iterations of the scales featured in this study could show stronger reliability and validity evidence, and could be combined successfully into a unified score.

Last, this study focused primarily on validity evidence based on internal structure and relations to other variables. While these factors are key components of a validity argument, additional evidence should be gathered in future studies. Future studies could also compare the validity estimates of a unified score to the validity evidence for each item type individually.

8.4. Conclusion

Overall, this study provides preliminary, yet promising evidence that innovative item types can be used to measure social and emotional skills in elementary school children. It supports the use of sentences over adjectives in developing items for this age group and provides concrete recommendations for future studies that can be done to further improve the current scales.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Dana Murano:Conceptualization, Formal analysis, Data curation, Writing - original draft.**Anastasiya A. Lipnevich**:Supervision, Writing - review & editing.**Kate E. Walton**:Conceptualization, Methodology, Writing - review & editing.**Jeremy Burrus**:Conceptualization, Methodology, Writing - review & editing.**Jason D. Way**: Methodology, Software, Writing - review & editing.**Cristina Anguiano-Carrasco**:Data curation, Conceptualization, Writing - review & editing.

Declaration of competing interest

None.

References

Abrahams, L., Pancorbo, G., Primi, R., Santos, D., Kyllonen, P., John, O. P., & DeFruyt, F. (2019). Social-emotional skill assessment in children and adolescents: Advances and challenges in personality, clinical, and educational contexts. *Psychological Assessment*. https://doi.org/10.1037/pas0000591 Advanced online publication.

ACT, I. (2018). *ACT Tessera technical bulletin.* Iowa City, IA: ACT, Inc.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs, 47*, i–171.

Anguino-Carrasco, C., Walton, K. E., Murano, D., Burrus, J., & Way, J. (2018). *Validity of ACT Tessera Unified Score.* Iowa City, IA: ACT Data Byte.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*(1-2), 9–30.

Berg, J., Nolan, E., Yoder, N., Osher, D., & Mart, A. (2019). Social-emotional competencies in context: Using social-emotional learning frameworks to build educators' understanding. Retrieved from https://measuringsel.casel.org/wp-content/uploads/2019/02/Frameworks-C.2-.pdf.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaire. *Psychological Methods, 18*, 36–52.

Burkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement,* 1–28. https://doi.org/10.1177/0013164419832063.

Burrus, J., & Brenneman, M. (2016). Psychosocial skills: Essential components of development and achievement in K-12. In A. Lipnevich, R. D. Roberts, & F. Preckel (Eds.). *Psychosocial skills and school systems in the 21st century: Theory, research, and practice*. Switzerland: Springer.

California Department of Education (2008). *California Healthy Kids Survey, student well-being in California, 2008–10: Statewide elementary results.* San Francisco, CA: WestEd Health and Human Development Program for the California Department of Education.

Campion, M. C., Ployhart, R. E., & MacKenzie, W. I. Jr. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. Human Performance, 27, 283–310.

Casillas, A., Robbins, S., Allen, J., Kuo, Y., Hanson, M. A., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology, 104*, 407–420.

Chernyshenko, O., Kankaraš, M., & Drasgow, F. (2018). *Social and emotional skills for student success and wellbeing: Conceptual framework for the OECD study on social and emotional skills. OECD education working papers, no. 17*Paris: OECD Publishing.

Denham, S. A. (2015). Assessment of SEL in educational contexts. In R. P. Weissberg, J. A. Durlak, C. E. Domotrovich, & T. P. Gullotta (Eds.). *Handbook of social and emotional learning: Research and practice* (pp. 285–300). New York, London: The Guilford Press.

Dueber, D. M., Love, A. M., Toland, M. D., & Turner, T. A. (2019). Comparison of single-response format and forced-choice format instruments using Thurstonian item response theory. *Educational and Psychological Measurement, 79*(1), 108–128.

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*(1), 405–432.

Fisher, P. A., Robie, C., Christiansen, N. D., Speer, A. B., & Schneider, L. (2019). Criterion-related validity of forced-choice personality measures: A cautionary note regarding Thurstonian IRT versus classical test theory scoring. *Personnel Assessment and Decisions, 5*, 49–61.

Halverson, C. F., Havill, V. L., Deal, J., Baker, S. R., Victor, J. B., Pavlpoulos, V., ... Wen, L. (2003). Personality structure as derived from parental ratings of free descriptions of children: The inventory of child individual differences. *Journal of Personality, 71*, 995–1026.

Heaven, P. C. L., Mak, A., Barry, J., & Ciarrochi, J. (2002). Personality and family influences on adolescent attitudes to school and self-related academic performance. *Personality and Individual Differences, 32*, 453–462.

Hooper, A. C., Cullen, M. J., & Sackett, P. R. (2006). Operational threats to the use of situational judgment tests: Faking, coaching, and retesting issues. In J. Weekley, & R. Ployhart (Eds.). *Situational judgment tests* (pp. 205–323). Mahwah, NJ: Lawrence Erlbaum.

Huebner, E. S. (1991). Initial development of the student's life satisfaction scale. *School Psychology International, 12*(3), 231–240.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*, 371–388.

John, O. P., & DeFruyt, F. (2015). *Education and social progress: Framework for the longitudinal study of social and emotional skills in cities.* Organization for Economic Cooperation and Development (EDU/CERI/CD(2015)13).

Kankaraš, M. (2017). *Personality matters: Relevance and assessment of personality characteristics. OECD education working papers, no. 157*Paris: OECD Publishing.

Kankaraš, M., Feron, E., & Renbarger, R. (2019). *Assessing students' social and emotional skills through triangulation of assessment methods.* Paris, France: OECD Education Working Papers NO. 208. Organization for Economic Cooperation and Development.

Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of Situational Judgment Tests (SJTs). *European Journal of Psychological Assessment, 32*(3), 230–240.

Kautz, T., Heckman, J. J., Diris, R., ter Weel, B. T., & Borghans, L. (2014). *Fostering and measuring skills: Improving noncognitive skills to promote lifetime success.* OECD, Educational and Social Progress.

Kyllonen, P. C., Lipnevich, A. A., Burrus, J., & Roberts, R. D. (2014). Personality, motivation, and college readiness: A prospectus for assessment and development. *ETS Research Report Series, 2014*, 1–48.

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology, 97*, 460–468.

Lipnevich, A. A., MacCann, C., & Roberts, R. D. (2013). Assessing noncognitive constructs in education: A review of traditional and innovative approaches. In D. H. Saklofske, C. B. Reynolds, & V. L. Schwean (Eds.). *Oxford handbook of child psychological assessment* (pp. 750–772). Cambridge, MA: Oxford University Press.

Marzano, R. J. (2015). Using formative assessment with SEL skills. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gullotta (Eds.). *Handbook of social and emotional learning: Research and practice* (pp. 336–347). New York, NY: Guilford.

McCrae, R. R., & Terracciano, A. (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology, 88*, 547–561.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531–552.

Mervielde, I., & De Fruyt, F. (1999). Construction of the hierarchical personality inventory for children (HiPIC). In I. Mervielde, (Ed.). *Personality psychology in Europe, proceedings of the eighth European conference on personality psychology*. Tilburg: Tilburg University Press.

Murano, D., Martin, J. E., Burrus, J., & Roberts, R. D. (2018). Feedback and noncognitive skills: From working hypotheses to theory-driven recommendations for practice. In A. A. Lipnevich, & J. K. Smith (Eds.). *The Cambridge handbook of instructional feedback*. Cambridge University Press.

OECD (2012), PISA 2009 technical report, PISA, OECD Publishing. http://dx.doi.org/https://doi.org/10.1787/9789264167872-en.

Osher, D., & Berg, J. (2017). *School climate and social and emotional learning: The integration of two approaches.* Pennsylvania State University: Edna Bennet Pierce Prevention Research Center.

Piaget, J. (1964). Development and learning. In R. E. Ripple, & V. N. Rockcastle (Eds.). *Piaget rediscovered* (pp. 7–20). New York, NY: Freeman and Company.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*, 322–338.

Primi, R., John, O. P., Santos, D., & De Fruyt, F. (2016). Development of an inventory assessing social and emotional skills in Brazilian youth. *European Journal of Psychological Assessment, 32*, 5–16.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*, 203–212.

Roberts, R. D., Martin, J. E., & Olaru, G. (2015). *A Rosetta stone for noncognitive skills: Understanding, assessing, and enhancing noncognitive skills in primary and secondary education.* Asia Society and Professional Examination Service.

Salgado, J. F., & Tauriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*, 3–30.

Schmitt, D., Allik, J., McCrae, R. R., & Benet-Martinez, V. (2007). The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38*, 173–212.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: An application to the problem of faking in personality assessment. *Applied Psychological Measurement, 29*, 184–201.

Tackett, J. L., Slobodskaya, H. R., Mar, R. A., Deal, J., Halverson, C. F., Jr., Baker, S. R., & Besevegis, E. (2012). The hierarchical structure of childhood personality in five countries: Continuity from early childhood to early adolescence. *Journal of Personality, 80*, 847–879.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analysis of fakeability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210.

Walton, K. E., Cherkasova, L., & Roberts, R. D. (2019). On the validity of forced choice scores derived from the Thurstonian item response theory model. *Assessment.* https://doi.org/10.1177/1073191119843585.

Wang, I., MacCann, C., Zhuang, X., Liu, O. L., & Roberts, R. D. (2009). Assessing teamwork and collaboration in high school students: A multi-method approach. *Canadian Journal of School Psychology, 24*, 108–124.

Ziegler, M., MacCann, C., & Roberts, R. D. (Eds.). (2012). *New perspectives on faking in personality assessment.* Chicago, IL: Oxford University Press.