PSYCHOMETRICS OF THE SSIS IN AN AUSTRALIAN POPULATION

A DISSERTATION

SUBMITTED TO THE FACULTY

OF

THE GRADUATE SCHOOL OF APPLIED AND PROFESSIONAL PSYCHOLOGY

OF

RUTGERS,

THE STATE UNIVERSITY OF NEW JERSEY

BY AMANDA SHERBOW

IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE

OF

DOCTOR OF PSYCHOLOGY

NEW BRUNSWICK, NEW JERSEY OCTOBER 2014

APPROVED:

_____
Ryan Kettler, PhD

_____
Nancy Fagley, PhD

DEAN:

_____
Stanley Messer, PhD

ABSTRACT

The Social Skills Improvement System (SSiS; Gresham & Elliott, 2008) is a multi-stage, broadband system for assessing children in Preschool through 12th grade. Two aspects of this system were analyzed in this study: (1) the Performance Screening Guides (PSGs), a brief criterion-referenced measure, and (2) the Rating Scales (SSiS-RS), a more extensive rating system. Two samples of Australian teachers in Bundamba (n = 15) and South Brisbane (n = 30) implemented the SSiS on elementary school students. This study has the following objectives: (1) Compare the psychometric properties of US and Australian samples and (2) determine if the Australian sample displays appropriate psychometric evidence to substantiate the use of such a tool in this population. Results indicated that internal consistency reliability was good for both samples across the domains and subdomains. For some of the domains, reliability was significantly higher in the Australian sample compared to the US sample. With respect to content validity, there were some differences with regard to which social skills each sample of teachers believed were important. Teachers in the Australian sample typically rated areas as being more important compared to the US sample. With regard to internal structure validity, correlations among PSGs and SSiS-RS domains were similar between the two different countries. In addition, conditional probability analyses indicated the PSGs work appropriately as the first stage of a multiple gating procedure. The confirmatory factor analyses indicated that the three-factor model had poor fit to the data in both samples. There were few differences between the factor loadings of the two samples. Lastly, in terms of validity evidence based on relations to other variables, the individual PSG scores demonstrated low to moderate ability to predict students who displayed difficulties on

corresponding areas of the Queensland Two Year Diagnostic Net, a monitoring system based on teacher judgment. Overall, the United States and Australian data demonstrated similar results for the areas explored. The results indicated adequate psychometric evidence for most of the areas measured.

ACKNOWLEDGEMENTS

This study would not have been possible without the support of many educators, researchers, and family members. I want to extend my gratitude to the children, parents, and educators who participated in the study, including everybody involved in the data collection procedures.

I would also like to thank a wonderful dissertation committee, including Drs Ryan Kettler and Nancy Fagley. I would like to thank Dr. Kettler for making this study possible, and supporting me throughout this process. I would like to thank Dr. Fagley for her availability and willingness to work through the complicated programming and statistics that this study required.

Finally, I would like to thank my family and friends who provided continual support throughout graduate school and the dissertation process. I am eternally grateful for my parents, Linda and Bruce Sherbow, who taught me about dedication and have always provided me never ending support. I'd also like to thank Jackie and Jessica, my big sisters, from whom I have learned countless skills that have helped me to grow and succeed. In addition, I'm grateful for the Greens for being my family away from home and for my friends who made this journey with me. Finally, I would like to thank E.J. who is my advocate, editor, and motivator. I would most certainly not be where I am now without you.

TABLE OF CONTENTS PAGE

LIST OF TABLES

Introduction

Social skills assessment and intervention are imperative activities within schools. The Social Skills Improvement System (SSiS; Gresham & Elliott, 2008) is a multi-stage, broadband system for assessing children in Preschool through 12th grade. The teacher form is helpful in assessing classroom behaviors and developing and implementing classroom interventions. Two aspects of this system were analyzed in this study: (1) the performance screening guides (PSGs), a brief criterion-referenced measure, and (2) the Rating Scales (SSiS-RS), a more extensive rating system. Although it has been shown to have adequate psychometric properties in the United States (US; Gresham & Elliott, 2008), little research has been done in other countries that may have interest in using the SSiS. Australian teachers implemented the SSiS with a sample of elementary school students. This study is designed to analyze the psychometric properties of the SSiS implemented in Australia and compare it to the previously examined data on a US sample to determine if the measure has precision with both.

Background and Problem

Social skills are a very important part of a child's development. Not only can social skills help promote relationships with others and future positive adjustment, social skills are related to other positive outcomes. Due to the numerous social interactions and situations that occur in a classroom, it is not surprising that social skills are related to academic achievement and school success. Academic achievement and competence have been shown to correlate positively with social skills (Malecki & Elliott, 2002; Welsh, Parke, Widaman, O'Neil, 2001). One study (Wentzel, 1993) examined the relationship between prosocial and antisocial classroom behavior (as measured by peer nominations

and teacher ratings) and academic achievement (as measured by grade point averages and standardized assessment). Results indicated that these kinds of social behaviors are significant, independent factors related to academic achievement, when other possible influencing variables (e.g., teacher's preference, IQ, family structure) were controlled in a multiple regression analysis.

While social skills have been linked to many desirable variables, social skills deficiencies have been linked to disadvantageous variables. Social skills deficiencies, or low social competence, are typical of many different types of emotional, developmental, and behavioral disorders including autism, specific learning disabilities, attention deficit/hyperactivity disorder, conduct disorder, and mild mental retardation (Gresham, Elliott, Vance & Cook, 2011). Lyon, Albertus, Birkinbine & Naibi (1996) found that disabled preschool students were rated as less skilled in several social skill areas, including cooperation, assertion, and self-control. Likewise, they were rated as showing a greater number of externalizing behavior problems. Exhibiting social difficulties in school has been linked with future maladjustment (Walthall, Konold, & Pianta, 2005). These relationships are observed throughout different stages of development. In fact, variables such as low self-esteem, loneliness, and lack of acceptance by others have been found to be negatively correlated with social skills in adolescence (Lyon et al. 1996). In adulthood, social skills difficulties have been linked to negative outcomes, such as mental health problems, poor adjustment, risk of depression, and risk of suicidal behaviors (Lyon et al. 1996).

Given this evidence, assessment focused on identifying students with social skills difficulties is particularly important in schools. In fact, Australian schools have recently

incorporated social-emotional learning criteria into their curriculum standards (Australian Curriculum, Assessment, and Reporting Authority [ACARA], 2011). Students are expected to reach a certain "Personal and Social Capability." This standard strives to guide students to recognize and regulate emotions, develop empathy for and understanding of others, establish positive relationships, make responsible decisions, work effectively in teams, and handle challenging situations constructively (ACARA, 2011). Much like any other area of the curriculum, it is important to have methods of assessment and intervention for students that are struggling. Without appropriate social skills assessment, interventions that target these skills may not be appropriate or effective (Merrell, 2001). Thus, there is a need for the development of assessment measures that accurately identify students with social skills difficulties in schools. Such measures must be validated in the populations in which they are used. This study focused on the reliability of the scores, and the validity of ensuing inferences from a social skills rating system developed in the US and used in an Australian population.

Defining and Measuring Behaviors

Social skills. Social skills and positive social behavior are adaptive behaviors that enable an individual to successfully navigate and interact in a social world. Social competence is a construct that is broader than social skills and refers to a child's adaptive functioning in his or her social environment (Rydell, Hagekull, & Bohlin, 1997). According to Gresham and Elliott (1987), social skills and adaptive behavior make up this general construct of social competence.

There have been two models discussed for conceptualizing the construct of "social skills"—the trait model and the molecular model (Gresham & Elliott, 1987). The

trait model posits that social skills are an underlying predisposition that is constant across all contexts and time. The molecular model suggests that social skills are situation specific behaviors that are not linked to an underlying personality characteristic or trait. The latter model is somewhat less abstract and is often used when assessing and measuring these skills in children due to the ease of operationally defining these skills/competencies in order to observe/identify them. The ways in which social skills are measured typically follow the molecular model, focusing on situation-specific behaviors. Social skills, for the purposes of this project, can be explicitly defined as "socially acceptable learned behaviors that enable a person to interact effectively with others and to avoid socially unacceptable responses" (Gresham & Elliott, 1990, p. 1).

There are many different methods to measure social skills in individuals (Lyon, Albertus, Birkinbine, & Naibi, 1996; Merrell, 2001). These methods include: behavior rating scales, naturalistic behavioral observation, interviewing, sociometric techniques, projective-expressive techniques, and objective self-report instruments. Merrell (2001) organized these methods into "first-line," "second-line," and "third-line" choices for social skills assessment. These designations indicate how often and how important they should be considered in social skills assessment, given their reliability and validity. "Third-line" methods include expressive-projective and self-report instruments, indicating that these methods are the least reliable and usable in the assessment of social skills. Sociometric techniques and interviewing are considered to be "second-line" because although they are appropriate to assess social skills in a given situation, they have many pragmatic problems.

Finally, direct behavioral observation and behavior rating scales are considered to be "first-line" assessment measures as there is much empirical evidence supporting their use in the assessment of social skills. Merrell (2001) described many advantages of using behavior rating scales. They are relatively inexpensive and provide data on low-frequency behaviors. These types of measures also provide relatively reliable data about individuals who may not be able to readily report on their own behaviors.

It is considered "best practice" to use a number of different data points in the assessment of social, emotional, and behavioral difficulties in students (Merrell, 2001). Behavior rating scales can be used as a multi-screening method and can identify students in the early stages of a behavioral, social, or emotional difficulty to identify students who might need more in-depth assessment (Merrell, 2001). One way in which behavior rating scales are often used is for multiple-phase or multiple-gating assessment. Multiple-gating is a sequential assessment procedure that involves universal screening before more specified assessment measures (Merrell, 2001). Through a series of decision steps a large population of interest is narrowed to a smaller, targeted population of those most at-risk. Universal screening is typically the first step. It can be defined as the "systematic assessment of all children within a given class, grade, school building, or school district, on academic and/or social-emotional indicators…" (Ikeda, Neessen, & Witt, 2009). Individuals who achieve a pre-determined cutoff score on a rating scale or other type of universal assessment measure are subsequently tested with more extensive measures and are typically referred for intervention, final classification, and/or even more involved assessment (Merrell, 2001).

As with any kind of psychological or educational assessment measure, a universal screener should be evaluated for reliability of its scores and validity of its inferences. A universal screening system should accurately predict students who are at-risk for certain difficulties and need intervention. The system must be able to distinguish between those who will and those who will not display difficulties in an area (Glovers & Alber, 2007). Due to the dichotomous nature of screening scores and inferences, conditional probability analyses can be used to evaluate this. Such analyses indicate the dichotomized level of agreement between risk status as defined by a universal screening system and actually having a condition, typically operationalized using a more extensive and established measure.

Relations among parts of an assessment measure, as well as relations with other variables, can be assessed another way to provide evidence of construct validity. That is, the evaluation of the matrix of intercorrelations among variables of representing at least two variables on at least two different methods can be assessed to provide evidence of construct validity. According to Campbell and Fiske (1959), measures of the same construct should produce higher correlations than with measures of different constructs. Furthermore, these correlations should be higher than correlations among different variables measured by the same method.

Problem behaviors. Much like social skills, problematic behaviors can be assessed using behavior rating scales. These scales have multiple purposes, including identifying children at risk for behavioral difficulties, organizing children into different levels of risk, and assessing the effectiveness of interventions (Cheong & Raudenbush, 2000). Across the literature, problem behaviors have been defined and conceived in

several different ways. Problem behaviors can be defined either through a general construct or through a more differentiated approach (Cheong & Raudenbush, 2000). Advocates of the former approach suggest that problem behaviors can be conceived as a single, general construct. The latter view allows for the identification of multiple, specific characteristics of problem behaviors.

The Child Behavior Checklist (CBCL; Achenbach, 1991) provides an example of how problem behaviors can be conceived. The CBCL is intended to measure two broadband, general factors—Internalizing and Externalizing behaviors. The internalizing behaviors factor of the CBCL, as well as throughout the literature, represents behaviors related to anxiety, depression, and inhibition. The externalizing behaviors factor represents behavior problems that are displayed in a child's outward behavior (Lieu, 2004). These problem behaviors typically involve a child's negative interaction with the external environment, such as disruptive or aggressive behavior. Scales of the CBCL are further subdivided into narrowband subscales to further differentiate the behaviors. Similarly, the Social Skills Improvement System, Rating Scales (SSiS-RS; Gresham & Elliott, 2008) contains a Problem Behaviors domain, as well as five subdomains measuring more specific constructs (e.g., externalizing, bullying, hyperactivity/inattention, internalizing, and autism spectrum).

Academic competence. Academic competence has been described in both broad and narrow terms within the literature (DiPerna & Elliott, 1999). DiPerna and Elliott (1999) considered achievement on standardized achievement tests as an example of how academic competence is conceptualized in narrow terms. A broad way to describe academic competence is as "a multidimensional construct composed of the skills,

attitudes, and behaviors of a learner that contribute to teachers' judgments of academic performance" (p. 208, DiPerna & Elliott, 1999). Areas that are involved in this construct include: study skills, interpersonal skills, academic motivation, academic self-concept, and academic skills. Teachers' judgments (e.g., ratings) of academic achievement have been shown to be related to academic performance (e.g., scores on achievement tests) across numerous studies (Hoge and Coladarci, 1989).

The Academic Competence Evaluation Scale (ACES; DiPerna & Elliott, 2000) is a teacher rating form that assesses students' academic competence in Kindergarten through college. Several areas of academic competence, as measured by the ACES, are positively related to Social Skills and negatively correlated with Problem Behaviors, as measured by the Social Skills Rating System (DiPerna & Elliott, 1999). The areas measured by the ACES include the following domains: Academic, Interpersonal, Motivation, Study, and Participation. The Interpersonal Relations domain on the ACES had the highest correlation with the Social Skills domain. The lowest correlation was with the Academic Skills domain. With regard to relationships with Problem Behaviors, the Interpersonal, Motivation, and Study scales demonstrated significant, negative correlations. The Academic Skills and Participation domains exhibited nonsignificant correlations, providing no evidence of linear relationships, with the Problem Behaviors domain.

Measurement standards. There are many factors to consider when evaluating the psychometrics of a behavior rating scale including: accessibility, reliability, construct validity, and consequential validity (Kettler & Feeney-Kettler, 2011). Of most interest to this study are reliability and construct validity. Reliability refers to the consistency of a

set of scores. Coefficient alpha is an indicator that characterizes the internal consistency of a test and is the average of all the possible split-halves of a set of scores (Cortina, 1993).

Construct validity refers to how well a test measures the construct that it is intended to measure. AERA, APA, and NCME (1999) described sources of validity evidence, but emphasized that these are not different types of validity. Instead, they are ways to find evidence for a validity argument for a particular test or measurement. Evidence can be garnered from a variety of sources including: test content, response processes, relations to other variables, internal structure, and consequences of testing (AERA, APA, & NCME, 1999). Several of these sources of evidence will be used in the current project, including test content, relations to other variables, and internal structure.

Teachers and assessment of social skills. Teachers are in an excellent position to rate children's behavior within school. They see the child for a large portion of the day and observe students in different contexts and situations. Gresham and Noell (1996) investigated the role of teachers as judges of students' social competence. Teachers (n = 208) used the Social Skills Rating Scales-Teacher Form (SSRS-T) on a sample of 1,021 students in K through 6th grade. Conditional probability methods were used to determine what items were probable in students rated in the low or high social skills groups. The results indicated that teachers' ratings of social skills function better at identifying adequate social competence than at identifying social incompetence. In addition, the items were more accurate predictors of low social competence for the female students. For the females, a total of 12 social skills were good predictors of both the presence and absence of social competence. The majority of these were on the cooperation subscale;

however, all subscales were represented. The results for the male students were similar to the results for the female students; however, fewer social skills were identified as accurately predicting both the presence and absence of social competence.

Teachers hold beliefs about what social skills are crucial for academic achievement and success in the classroom. These views play an important role in social skills assessment and intervention. One study investigated elementary school teachers' (*n* = 126) expectations of student behavior and social skills that were critical for success in their classroom, as measured by the SSRS-T (Lane, Givner, & Pierson, 2004). The study also addressed whether or not there were any differences between which social skills special education teachers and general education teachers deemed to be most important. Teachers rated self-control and cooperation as being equally important. Assertion skills were rated as less important. Both general education teachers and special education teachers rated self-control skills and assertion skills as equally important; however, general education teachers viewed cooperation skills as more essential for success in the classroom (Lane et al., 2004).

Meier, DiPerna, and Oster (2006) replicated and extended the Lane et al. (2004) study. Teachers' judgments of the importance of social skills were investigated for first through sixth grade teachers. This study also explored the stability of teachers' ratings over time. Elementary school teachers (*n* = 50) completed the importance rating section of the social skills subscale of the SSRS-T at the beginning and at the end of the school year. Paired t-tests were used to determine what social skill domains (i.e., Cooperation, Self-Control, and Assertion) teachers view as most important for success in the classroom and to determine if teachers' ratings significantly changed from the beginning of the year

to the end of the year. Results indicated that cooperation and self-control were both more important than assertion. In addition, self-control was rated significantly higher than cooperation. Finally, the teachers' ratings of importance were stable from time 1 to time 2 for each of the domains.

Similarly, in the standardization sample of the SSiS-RS, teachers rated that many of these same factors are critical in the classroom setting (Gresham & Elliott, 2008). The following skills were rated by teachers as the most critical skills to have in the classroom: (1) listens to others, (2) follows directions, (3) follows classroom rules, (4) ignores peer distractions, (5) asks for help, (6) takes turns in conversations, (7) cooperates with others, (8) controls temper in conflict situations, (9) acts responsibly with others, and (10) shows kindness to others. The items on the cooperation subscale were perceived as most important. The authors suggest that this may be due to the perception that when students cooperate, the classroom is less disruptive. A rating system for social skills should target skills that are important to the child's development, as well as skills that are important to the teacher in terms of constructing interventions to improve classroom success. Lane et al. (2004) suggest that if interventions do not result in changes in social skills that the teacher deems important for classroom success, intervention fidelity may suffer and more extensive, and restrictive, interventions may instead be implemented.

Psychometrics of US-Developed Measures Implemented in Other Populations

Although the most researched behavior rating scales focusing on social skills have been tools initially developed in the US, it is common for their psychometrics to be investigated in other countries (Matson & Wilkin, 2009). Research regarding the

psychometric properties of US-normed behavior rating scales has shown important results in other countries.

The Children's Self-Report Social Skills Scale. The Children's Self-Report Social Skills Scale (CS; Danielson and Phelps, 2003), a 21-item self-report rating scale, was translated into Turkish and implemented on 4th, 5th, and 6th grade students attending elementary schools in Turkey (Gendongan, 2008). To examine the internal structure validity evidence of the translated measure, an exploratory principal components analysis was conducted. Results were comparable to the factor analysis conducted on the English version analyzed with a sample of US students (Danielson and Phelps, 2003). The three-factor model of the scales was confirmed. The three factors include Social Rules, Likeability, and Social Ingenuousness. This is congruent with the studies conducted in the US, as well as the way in which the construct of social skills was conceptualized to guide the development of the tool. Construct validity, as evidenced by relations to other measures, was evaluated by comparing the scores on this measure to similar and diverging measures. Results were as expected in that high social skills on the CS were negatively related to hopelessness and scores on the CS were positively related to peer nomination scores. The authors suggest that these results support the use of this measure in Turkey in its translated form.

The Strengths and Difficulties Questionnaire. The Strengths and Difficulties Questionnaire (SDQ; Goodman, 2001) is used to assess youth ages 3 through 16 based on 25 items related to positive and negative characteristics, using a 3-point Likert scale (0 = *Not True*, 1 = *Somewhat True*, 2 = *Certainly True*). Hawes and Dadds (2004) examined the psychometric properties of the SDQ in an Australian sample (n = 1359).

Psychometric properties studied included: reliability of the scores and construct validity in terms of relations with other variables and internal structure. Reliability, as measured by Cronbach's alpha (Cronbach, 1951), was found to be low to moderately high (ranging from .59 to .82) across all subscales. In addition, the original five-factor structure of the measure was generally confirmed via principal components analysis. Test-retest reliability was examined, and it was found that for a 12-month period, parents' ratings were adequately stable. These correlations ranged from .61 to .77, which were only slightly lower than the correlations that resulted from a shorter test-retest interval (i.e., 1-2 months). The authors concluded that these correlations provide adequate evidence of stability given the prolonged test-retest period. To provide evidence of construct validity with regard to relations to other variables, scores on the SDQ were compared to clinician perceptions of Axis 1 diagnostic criteria. Conduct problems, hyperactivity, and emotional symptoms resulted in significant, albeit medium-sized correlations, with the clinical assessments of externalizing disorders, with correlations ranging from .33 to .51. Further, peer problems as measured by the SDQ were positively related to diagnoses of conduct problems and internalizing disorders. In addition, an ANOVA analysis was conducted, and it was found that those receiving treatment for behavioral or emotional difficulties, according to parent reports, received a higher SDQ overall score compared to students not receiving treatment.

ADHD Rating Scale IV. Zhang, Faries, Vowles, and Michelson (2005) investigated the reliability of the scores and validity evidence of the ADHD Rating Scale IV (DuPaul, Power, Anastopoulos, & Reid, 1998) in 11 European countries, as well as Australia, Israel, and South Africa. The authors examined the following psychometric

evidence and found positive results: interrater reliability, test-retest reliability, and construct validity (internal structure and relations to other variables). A total of 604 patients with attention deficit hyperactivity disorder (ADHD) were rated using the scales. The patients were rated during two different visits, about 1 week apart. Interrater reliability, as measured by Kappa, was satisfactory and similar to results in a US trial. Cronbach's alpha (Chronbach, 1951) was measured to be .80 at Visit 1 and .84 at Visit 2, providing evidence of the measure's internal consistency. A factor analysis revealed that the internal structure of the scales across the different countries fit a 2-factor model, consistent with the hypothetical and theoretical constructs related to the diagnosis of ADHD that incorporate inattention and hyperactivity as factors. Validity with regard to relations to other variables of the scale was described as adequate due to its high correlations with other ADHD measures and its low correlations with areas that were thought to measure different symptoms. Overall, the results from this study supported the validity and reliability of this rating scale to aid in the diagnosis of ADHD in other countries.

The Social Skills Rating System. A translated version of the Social Skills Rating System (SSRS: Gresham & Elliott, 1990) was used with preschool children in Iran to determine the reliability of SSRS scores in this population (Shahim, 2004). The purpose was to investigate the stability and internal consistency of the SSRS in a group of students aged three to six and a half (n = 304). Results indicated that Cronbach's alpha (Chronbach, 1951) coefficients were .71 for the Social Skills domain and .69 for the Problem Behaviors domain, indicating low to moderate reliability. These values were lower than the reliabilities resulting from standardization data from the US and indicate

that the SSRS scores are moderately stable in this sample. The authors noted that further evidence should be collected to investigate the construct validity in this population.

The Social Skills Improvement System. Kettler, Elliott, Davies, and Griffin (2011) investigated the predictive validity of the SSiS-RS with a sample of Australian Elementary school students (n = 360). In this study, the criterion was end-of-year achivement and the predictor was scores on the SSiS teacher Rating Scales and the PSGs. The goal was to determine the optimal combination of scores that would best predict academic difficulties. First, reliability of scores within the sample was estimated. The Australian sample was found to have similar reliability estimates as the US sample for the Rating Scales. Correlations among subscales on the PSGs and on the Rating Scales were calculated and provided evidence for internal structure validity. Through conditional probabilitity and multiple regression analyses, it was concluded that the PSGs and the Rating Scales are both predictive of academic achievment and work best when used together in this Australian sample. This study provided important information on the SSiS implemented in Australia. However, several psychometric aspects of this sample have yet to be studied.

Summary

Social skills are an important part of development and are associated with positive outcomes. The SSiS, a behavior rating system, was developed to assess social skills in children as measured by teachers, parents, and the students themselves. Teachers play an important role in the assessment of social skills. Teacher ratings of these skills can be reliable and accurate in predicting social difficulties in their students. Furthermore, teachers hold perceptions with regard to what social skills are most important in the

classroom. Several studies have evaluated the use of behavior rating systems, including the SSRS and the SSiS, with other populations outside the normative population. These studies have looked at different types of reliability (e.g., test-retest, internal consistency) and validity (e.g., internal structure validity and relations with respect to other variables) of a number of behavior rating scales in different locations. The current study is comparable to these studies, while also extending the research by providing additional information on a social skills behavior rating system implemented in Australia.

Current research questions

The current study utilized extant data from the SSiS implemented in Australian samples of elementary school children. The objective of this study was to assess the psychometric properties of the SSiS in the Australian samples. The psychometrics of the US standardization and the Australian samples were compared when possible. This investigation was designed to contribute to existing literature regarding the SSiS by allowing parallel analyses and direct comparison of the instrument system implemented in samples from two different countries. Additionally, it was aimed at extending the previous work utilizing this data (i.e., Kettler et al., 2011) in that different types of psychometric analyses will be conducted. Overall, the study was intended to provide additional evidence on whether the SSiS is psychometrically sound and suitable to be used in an Australian population. The study aimed to address the following research questions:

1. Are there differences in the reliability of SSiS-RS scores between Australian and US samples? The reliability estimates of the US and Australian population were predicted to be comparable based on previous studies depicting moderate

to high estimated reliability coefficients for the SSiS/SSRS conducted in various populations.

2. Are there differences in content validity evidence for SSiS-RS scores between Australian and US samples? It was predicted that both populations were comparable with regard to what social skills are believed to be most important.
3. Are there differences in internal structure validity evidence for the SSiS between Australian and US samples? Internal structure validity was predicted to be consistent between the two populations given the theoretical and evidence-based construction of the scales.
4. Are there mean differences in SSiS-RS and PSG scores between the Australian and US samples? Mean scores were predicted to be similar between the two samples.
5. Do the relations between SSiS scores and the Year Two Diagnostic Net, a teacher monitoring system for academic difficulties, provide evidence for construct validity with respect to relations with other variables? The SSiS academic areas were hypothesized to accurately predict students' scores on the Year Two Diagnostic Net.

Method

Participants

Three samples of data were used for the analyses in this study. The first two samples were data collected in classrooms in Queensland, Australia (i.e., in South Brisbane and in Bundamba) and the third is the US Standardization sample. Table 1

identifies the samples used for this study and details how they relate to the subsequently described analyses. The table also identifies previous studies using the Brisbane South District data and the US SSiS standardization data and depicts how the previously conducted analyses compare to the current analyses.

Table 1
Samples Paired With Analyses

	Current Study						Previous Studies			
	Brisbane South District		Bundamba		United States		Kettler, Elliott, Davies, & Griffin (2011)		SSIS Manual	
	Total (*n* = 536)		Total (*n* = 321)		Teacher Ratings (*n* = 950)		Brisbane South District grades 3 & 5 (*n* = 360)		Teacher Ratings (*n* = 950)	
	PSG (*n* = 536)	RS (*n* = 179)	PSG (*n*= 321)	RS (*n* = 101)	PSG (*n* = 85)	RS (*n* = 950)	PSG (*n* = 360)	RS (*n* = 178)	PSG (*n* = 85)	RS (*n* = 950)
Internal Consistency Reliability		X		X		X		X		X
Content validity		X				X				
Internal structure validity: confirmatory factor analysis		X		X		X				
Internal structure validity: correlations among subscales	X	X	X	X	X	X	X	X		X
Internal structure validity: conditional probability analyses	X	X			X	X				
Validity evidence based on external factors: comparing the PSG and NET	X	X								
Means Calculated	X	X	X		X	X	X	X	X	X

The first sample consisted of 30 teachers in the Brisbane South district in Queensland, Australia who used the SSiS (Elliott & Gresham, 2008) in their second, third, and fifth grade classrooms. The Brisbane South district contains 36 primary schools across a range of SES areas. A total of 536 students were rated using the PSGs. A stratified (i.e., stratified on gender and levels on the PSGs) random subsample of this sample (n =179) was selected to be rated by their teachers using the Rating Scales. The mean ages in months of those rated on both the PSGs and the Rating Scales and of the

subset rated on the Ratings Scales were 96.9 (*SD* = 14.9) and 97.7 (*SD* = 15.4), respectively.

Another Australian sample was used for these analyses. This sample includes 321 students in Prep (first year of schooling—Kindergarten) through Year 3. In this sample, fifteen teachers utilized the PSGs to evaluate the students in their classes. Those students who were judged at level 1 or 2 (at-risk) on the prosocial performance area were further assessed using the SSiS-RS, teacher form (*n* = 101). The mean age in months was 77.2 (*SD* = 14.2) for the PSGs and 73.9 (*SD* = 8.1) for the subset also rated on the Rating Scales. Ethnicity and SES are not typically documented in Australia. Therefore, this information is not available for the Australian samples. See Table 2 for demographic information concerning the Australian samples.

Table 2
Demographic Information for the Australian Samples

	Brisbane South District (*n* = 536)		Bundamba (*n* = 321)		Total (*n* = 857)	
	PSG (*n* = 536)	RS (*n* = 179)	PSG (*n*= 321)	RS (*n* = 101)	PSG (*n* = 857)	RS (*n* = 280)
Gender (%)						
Female	50	49	47	40	49	46
Male	50	51	53	60	51	54
Grade (%)						
Prep	0	0	28	41	11	15
1	0	0	28	31	11	11
2	33	34	21	7	29	24
3	40	37	22	22	33	31
4	0	0	0	0	0	0
5	27	30	0	0	17	19

In addition, data collected for the US standardization sample for the SSiS-RS, Teacher Form (*n* = 950) and the PSGs (*n* = 85) was used for the comparative analyses. Data were collected as part of the national norm study.

The mean age in months was 112.2 (*SD* = 49.0) and 93.5 (*SD* = 40.40) for the US sample. Additional demographic information representing this sample is located in Table 3.

Table 3
Demographic Information United States Sample (*n* = 950)

	Rating Scales (*n*= 950)	PSG (*n* = 85)
	%	*%*
Gender		
Female	50	38
Male	50	62
Ethnicity		
Black	16	20
Hispanic	20	7
White	58	66
Other	6	7
Grade		
Daycare	.3	0
Preschool	21	26
K	9	5
1	9	23
2	8	5
3	8	6
4	6	5
5	8	5
6	7	5
7	7	14
8	6	7
9	3	0
10	3	0
11	4	0
12	3	0

Instruments

The Social Skills Improvement System (SSiS). The SSiS (Gresham & Elliott, 2008) is a revised version of the Social Skills Rating System (Gresham & Elliott, 1990). The measures are comparable with regard to psychometrics, exhibiting high internal consistency estimates and moderately high validity indices (Gresham, Elliott, Vance, &

Cook, 2011). The SSiS is a multi-tiered, comprehensive assessment and intervention system that addresses important social skills in students from age 3 to 19. The system includes a class-wide, universal screening tool (the Performance Screening Guides), a more targeted assessment instrument for at-risk students (the SSiS Rating Scales), and intervention guides to help plan interventions based on the results of the Performance Screening Guides (PSGs) and the SSiS Rating Scales (SSiS-RS).

Performance Screening Guides (PSGs). The PSGs (Gresham & Elliott, 2008) are tools designed for class-wide screening of students' social, motivational, and academic skills. Students' abilities in a given skill area (i.e., Prosocial Behaviors, Motivation to Learn, Reading Skills and Mathematics Skills) are measured against grade-level expectancies. The PSGs are designed for students aged 3 to 17 years old. During the administration of this tool, teachers are asked to choose the performance level that best represents each of their students' current levels of functioning in a given performance area based on several weeks of observations and interactions. Students earn performance ratings of 1 through 5 in each of the aforementioned four skill areas. Ratings are grouped into color bands, which represent the amount of intervention necessary for that particular skill area. A performance rating of 1 (red band) indicates an area in need of direct and remedial instructional action. Performance ratings of 2 or 3 (yellow band) indicate "caution" and require teacher attention to monitor the students' functioning in that particular area. Students earning a performance rating of 4 or 5 (green band) in a skill area are identified as not at-risk and additional attention is not necessary for that skill area. This process typically takes about 25 minutes per classroom.

Concerning psychometric properties, the authors note that reliability measurements were adequate, given the brevity of the forms (Gresham & Elliott, 2008). Test-retest reliabilities ranged from .68 to .74 for the Elementary (n = 302) and Secondary (n = 177) levels and inter-observer reliabilities ranged from .55 to .68 across skill areas for the Elementary (n = 215) and Secondary (n = 140) levels. The validity study compared individuals rated on both the PSG and the SSiS-RS. The correlation between Prosocial PSG scores and Social Skills domain scores was r = .70 for the Elementary/Secondary levels (n = 63). All correlations with the Problem Behavior scales were moderate and negative.

The SSiS Rating Scales (SSiS-RS). The SSiS-RS (Gresham & Elliott, 2008) is used for targeted assessment of students' social skills, problem behaviors, and academic competence. Teacher, parent, and self-report forms allow for comprehensive assessment of behaviors across contexts for children ages 3-18. The teacher form, which consists of 82 items, was used in the current study. There are three main domains for which scale scores are calculated: Social Skills, Problem Behaviors, and Academic Competence. Within the Social Skills domain, there are seven subscales: Communication, Cooperation, Assertion, Responsibility, Empathy, Engagement, and Self-Control. The frequency of each item is rated using a 4-point frequency scale (0 = *Never,* 1= *Seldom,* 2 = *Often* and 3 = *Almost Always*). In addition to the frequency scale, the importance of each item is rated using a 3-point scale (0 = *Not Important,* 1 = *Important,* 2 = *Critical).*

The Problem Behaviors scale is composed of the following subscales: Externalizing Behaviors, Internalizing Behaviors, Hyperactivity/Inattention, Autism Spectrum, and Bullying. The aforementioned frequency rating scale is used to obtain a

score on the Problem Behaviors domain.

For the Academic Competence scale, teachers rate the students' levels of academic competence in Reading Performance, Mathematics Performance, and Motivation to Learn relative to the entire classroom using a 5-point scale (1= *Lowest 10 % of the class*, 2 = *next lowest 20%*, 3 = *the middle 40%*, 4 = *the next highest 20%*, 5 = *Top 10% of the class*).

For each of the three domains (Total Social Skills, Total Problem Behaviors, and Total Academic Competence), a norm-referenced scaled score is calculated (M = 100; *SD = 15).* Scores within one standard deviation of the mean are considered to be average. Scores below 85 are considered below average and scaled scores above 115 are considered to be above average. Subscale scores for social skills and problem behaviors are categorized into "behavior levels," which are descriptive characterizations (e.g., "below average," "average," and "above average") of the scores in comparison with the normative group.

Internal consistency reliability, test-retest reliability, and interrater reliability were calculated using the normative sample of 4,700 children and adolescents. Evidence of internal consistency reliability shows coefficient alphas in the upper .90s for the Social Skills, Problem Behaviors, and Academic Competence domains. Subscale reliabilities were in the .80s for teacher forms. Test-retest reliability indices on the teacher forms were .82 for the Social Skills domain and .83 for Problem Behavior domain. Test-retest reliabilities for the Academic Competence domain were even higher. The time interval for the test-retest reliability was an average of 43 days for the teacher form. In terms of validity, correlations among social skills subscales were positive and moderate to high, as

were expected. The relationships between Social Skills and Problem Behavior scales were moderate and negative for all forms. For evidence of concurrent validity, the SSiS-RS was compared to other relevant scales, such as the SSRS (Gresham & Elliott, 1990), The Behavior Assessment Scale for Children (Reynolds & Kamphaus, 1998), and The Vineland Adaptive Behavior Scales (Sparrow, Balla, Cicchetti, 2005). Results from these comparison analyses provide evidence of the concurrent validity of the SSiS-RS.

There have been limited investigations into the factor structure of the SSiS-Rating Scales. The SSiS manual indicates that factor analysis was conducted and utilized when constructing the Social Skills scale, but the results were not thoroughly described. According to the manual, the seven-factor structure of the Social Skills Scale was a modest overall fit to the data as expressed by Comparative Fit Index (values in the mid-.80s). However, this analysis was done only to identify possible beneficial changes during the construction of the tool, rather than for the purpose of testing a factor model. Despite the lack of evidence regarding the factor structure of the SSiS, there has been a study regarding the factor structure of the SSRS. Walthall, Konold & Pinata (2005) investigated the factor structure of the SSRS across gender and ethnicity in a sample of 995 elementary school students. They were interested in investigating the three-factor structure of this instrument and whether or not it differed based on student ethnicity and gender. Results indicated that a three-factor model, as described in the SSRS Teacher form manual, provided a solid representation for the data collected for this study as indicated by a range of goodness of fit measures. Multi-group confirmatory factor

analysis addressed the issue of invariance based on gender and ethnicity. Results indicated that the three-factor model was supported across gender and ethnicity categories.

The Year Two Diagnostic Net. This monitoring system was developed by the Queensland Studies Authority to support the development of literacy and numeracy in the first few years of school (Queensland Government: Department of Education, Training and Employment, 1998). Students in the first three years of school are monitored to identify those who need additional support in the classroom. Teachers monitor students' progress in Reading, Writing, and Numeracy using *key indicators*, identified milestones of literacy and numeracy development. During this process, standard written reports are created and given to parents to inform them of their child's progress. A child receives a rating of A on the low end, B, C, D, or E on the high end in mathematics, reading, and writing. These ratings represent phases of development or competence levels increasing from the low rating of A to the highest rating of E. An average rating for a student is C and additional support is not required. If a child is considered to be in phase A or B, more targeted investigations are used to assist the child in the classroom. In this study, the Year Two Diagnostic Net was used as a criterion variable.

Procedure

Analyses on the South Brisbane Australian sample were conducted using an extant database collected by Kettler et al. (2011). Principals of 12 schools were contacted to seek involvement of their schools and their teachers. Nine of these principals agreed and informed their teachers about the opportunity to participate in the study. Teachers attended two half-day workshops that provided an overview of the research and the tools

that would be used. The teachers who agreed to participate sent information sheets and consent forms to the parents of all their students. Seventy-six percent of parents consented to have their children participate in the study. All of these students were rated on the PSGs by their teachers. Then, teachers rated a subsample of these students, obtained via stratified random sampling, on the Rating Scales.

An additional Australian sample was obtained from the Bundamba project, which involves the comprehensive and longitudinal implementation of a school-wide intervention. This project is aimed at examining the use of the SSiS in classrooms. Fifteen teachers used the PSGs to evaluate the students in their classes. Students who were rated to be at levels 1 or 2 on the Prosocial behavior domain were further assessed using the SSiS-RS teacher rating form. The assessments of these students were used to inform the teachers of the specific social skills to be targeted using the Class-wide Intervention Program.

As for the US standardization data, information provided in the examiner's manual was used for several of the analyses, along with additional data requested for the purposes of this study. For this standardization sample, children and adolescents (n = 4,700) were assessed using the three different forms of the Rating Scales at 115 different sites. Site coordinators recruited teachers to participate in the study. Participating teachers sent consent forms home, along with the SSiS-RS, a description of the project, and a form asking for demographic information. It was requested that each participating teacher complete no more than 6 of the forms. A subsample of these teachers was randomly drawn to complete the PSGs for their students.

Plan of Analysis

The analysis of the data had five main objectives, each of which is intended to compare the psychometrics of the Australian and US samples.

Reliability. Reliability, the consistency of a measure, was estimated using Cronbach's alpha (Cronbach, 1951), or coefficient alpha, at the scale and subscale level for the Rating Scales in Australian and US samples. To describe and interpret the reliability coefficients, Murphy & Davidshofer (2005)'s categorical descriptions were utilized. According to this system, coefficients ranging from .00 to .59 indicate very low reliability. Values ranging from .60 to .69 are considered to signify low reliability. Reliability coefficients ranging from .70 to .79 indicate moderate reliability. Values ranging from .80 to .89 represent moderately high reliability. Finally, values ranging from .90 to .99 imply high reliability. Coefficient alpha values were compared to the reliability estimates from the US standardization sample using the independent samples Feldt test (Feldt, 1969).

Construct validity. Several sources of validity evidence were analyzed in the current study, including evidence from test content, internal structure, and relations to other variables.

Content validity. Content validity refers to the degree to which the content of a measure appears to represent the construct that it was designed and intended to measure (AERA, APA, & NCME, 1999). Reviewing the literature, obtaining expert review, and field-testing are typical ways of gathering this type of evidence. To determine if the SSiS-RS item content represents which social skills are critical in the Australian population, the importance ratings of the Social Skills domain were analyzed. This facilitated in

determining if the Social Skills domain targets defining features of critical social skills in this sample. The mean importance ratings of the Social Skills domain of the SSiS-RS were compared via independent samples t-tests at the subscale level between the US and the Australian samples. These results provided evidence indicating whether or not the samples are similar with regard to what skills are crucial within a classroom setting.

Internal structure validity. To explore the internal structure of the SSiS, correlations between all scales and subscales on the PSGs and the Ratings Scales were calculated for the Australian and US samples using Pearson product-moment correlations. Hopkin's (2001) guidelines for classifying correlations were utilized to categorize and interpret correlations. These criteria are an extension of Cohen's (1992) guidelines. The classification system includes the following categories: nonexistent (between $r = .00$ to $r = .10$), small (between $r = .10$ and $r = .30$), medium (between $r = .30$ and $r = .50$), large (between $r = .50$ and $r = .70$), very large (between $r = .70$ and $r = .90$), and nearly perfect ($r = .90$ and above). Note that the same system was used for negative correlations.

In addition, confirmatory factor analysis was utilized to further examine evidence of internal structure validity. SPSS Amos was used to compute the confirmatory factor analysis. These results indicated whether the data-based pattern of relations among items matched the intended theoretical model depicted in the organization of the SSiS-RS for the Australian and the US samples. Because the SSiS-RS was designed based on pre-determined theory and evidence, confirmatory factor analysis is appropriate for examining internal structure validity. The structure of the SSiS-RS is composed of three scales (social skills, problem behaviors, and academic competence), each of which

includes at least four subscales. This analysis determined if the US and Australian data appropriately fit within this intended structure.

Goodness of fit indexes were computed in order to indicate how well the specific data is structured in relation to the proposed model. It is beneficial to calculate multiple indicators of goodness of fit when evaluating fit of a model (Thompson & Daniel, 1996; Myers, Gamst, & Guarino, 2006). The following indicators were calculated: The goodness-of-fit index (GFI), the comparative fit index (CFI), normed fit index (NFI), and the root mean square error of approximation (RMSEA). The GFI, an absolute fit measure, measures the amount of the variance in the sample covariance/correlations that is explained by the predicted model, and a statistic of .90 or higher is indicative of an adequate model (Meyers, Gamst & Guarino, 2006; Joreskog and Sorbom, 1986). The CFI measures the fit relative to an independence model. A statistic of .90 or higher likewise designates good fit (Bentler, 1990). The NFI, which indicates how well the proposed model improves the fit compared to the independent model, should have a value of .95 or higher for it to be deemed acceptable (Meyers, Gamst, & Guarino, 2006). The RMSEA evaluates the degree to which a proposed model fits reasonably well in the population (Brown, 2006). An indication of good fit is a statistic less than .08 (Loehlin, 2004). The current study will compare the relative fit of a Three First Order Factor (3F) model, consisting of all three domains as depicted in the manual and scoring system of the SSiS-RS.

Additional evidence of internal structure validity was obtained using conditional probability analyses. The SSiS is designed for use as a multi-tiered assessment procedure. As such, it is assumed that the PSGs should accurately distinguish students who have

social skills difficulties, and thus need more targeted assessment, from those who do not. If the SSiS is to be used as a multiple gating procedure, scores on the PSGs should predict scores on the Rating Scales. To date, no study has targeted this question. Conditional probability indices are often used to evaluate the validity of a screening system based on this area of construct validity (Kettler & Feeney-Kettler, 2011). This type of analysis is used when types of data can be organized dichotomously (e.g., at risk or not at risk). Thus, conditional probability analyses were utilized to determine and compare how well scores on the PSGs predict scores on the SSiS-RS for both the US standardization sample and the South Brisbane Australian sample. Students who received ratings of 3 or lower (yellow band or red band) in any of the areas on the PSG were considered at-risk for social, behavioral, and academic difficulties. Students above the average range for behavior problems or below the average range for social skills or academic competence on the SSiS-RS were designated as having problematic skills and/or behaviors.

Relations to other variables. Another type of validity evidence is based on relations to other variables (AERA, APA, & NCME, 1999). Information gathered for this type of validity evidence specify how scores on a screening system: (1) converge with scores from assessments that aim to measure similar constructs, (2) diverge on measures that assess opposing constructs, and (3) are distinct from scores on unrelated constructs (Kettler & Feeney-Kettler, 2011). Evidence of validity based on relations with other variables was assessed by conditional probability analyses between the PSGs scores of second graders and the Year Two Diagnostic Net (hereafter Net), a monitoring system based on teacher judgment. This determined whether the PSGs are adequate predictors of

academic difficulties, as measured by scores on the Net. The Academic Competence domain of the SSiS-RS was also used in these analyses to determine how well both measures together predict scores on the Net. Ratings in the yellow or red band on the PSGs in any of the four behavior domains indicated that a student was at-risk. Students with a "B" or an "A" on any of the academic areas of the Net were considered to have an academic difficulty when conducting these analyses.

Sensitivity, specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) was calculated for the conditional probability analyses. Sensitivity refers to the probability that a screening measure will accurately identify a child with an academic/behavioral difficulty (i.e., a valid positive). Specificity is the likelihood that a screening test will correctly not identify a child who does not have an academic/behavior/social difficulty (i.e., a valid negative). Positive Predictive Value specifies the likelihood that a student who was identified by the screening system does in fact have a difficulty. Finally, Negative Predictive Value specifies the probability that a student who is not identified by a screening system is a student who does not have a difficulty. Criteria for judging acceptable values on these indices may differ depending on the prevalence of a difficulty in a population and the relative costs of a false negative versus a false positive (Kettler & Feeney-Kettler, 2011). Indices between .80 and 1.00 on Sensitivity, Specificity, PPV, and NPV are considered to be in the high range (Kettler & Feeney-Kettler, 2011). Glover and Albers (2007) suggest that the usability of a screening system should be questioned when values fall below .75 or .80.

Mean differences. Independent samples t-tests were conducted to compare the frequency rating means between the Australian samples and the US standardization

sample on the SSiS-RS teacher form and on the PSGs. Comparisons were made between the samples' mean scale scores and subscale scores to determine whether Australian and US raters identified different degrees of concerns about their students' social skills, academic skills, and problem behaviors.

Analysis of this data has been approved by the Institutional Review Board of Rutgers, The State University of New Jersey.

Results

Reliability

The first research question aimed at examining whether the internal consistency of the SSiS-RS differed between the US and the Australian samples. To address this question, coefficient alpha for the SSiS-RS was calculated for the domain and subdomain levels for both samples (see Table 4). Coefficient alpha was in the high range for all domains (Social Skills, Problem Behaviors, and Academic Competence) in both countries. For the Social Skills domain, the alpha coefficient was .973 and .967 for the Australia and US samples, respectively. For the Problem Behaviors domain, coefficient alpha was .954 and .949 for the Australian and US samples, respectively. Similarly, the Academic Competence domain had reliability estimates of .974 and .964 for the Australian and US samples, respectively.

Table 4

Comparison of Coefficient Alpha of Australian and US Samples

SSiS Scales	Australian Sample (*n* = 280)	United States Sample (*n* = 950)	*p*
Social Skills	**.973**	**.967**	**0.021***
Communication	.905	.858	0.000*
Cooperation	.944	.908	0.000*
Assertion	.833	.849	0.858
Responsibility	.909	.905	0.335
Empathy	.932	.909	0.002*
Engagement	.905	.860	0.000*
Self-Control	.930	.906	0.002*
Problem Behaviors	**.954**	**.949**	**0.148**
Externalizing	.949	.934	0.005*
Bullying	.903	.866	0.001*
Hyperactivity/inattention	.911	.901	0.141
Internalizing	.845	.803	0.008*
Autism Spectrum	.908	.882	0.006*
Academic Competence	**.974**	**.964**	**0.000***

Note. The Independent Samples Feldt Test was used to determine significant differences. For the Academic Competence scale for the US sample, *n* = 744 because preschool students were not rated on this area.
*p < .05

Regarding the subdomain areas, the Australia sample had alpha coefficients in the moderately high and high range. The subdomain areas of Assertion (part of the Social Skills domain) and Internalizing (part of the Problem Behaviors domain) resulted in coefficients of .833 and .845, respectively. All other subdomain alpha coefficients ranged from .90 to .95, indicating high reliability. The subdomain alpha coefficients for the US sample were within the moderately high to the high range.

To assess the equality of the reliability coefficients of the US and the Australian samples, the independent samples Feldt test (Feldt, 1969) was computed. Results indicated that the Australian sample produced larger coefficients for the Social Skills and Academic domains, as well as for several subdomains (i.e., Communication, Cooperation, Empathy, Engagement, Self-Control, Externalizing, Bullying, Internalizing, and Autism spectrum) compared to the US sample. The largest raw difference in coefficient alpha values between the two samples was .047, which resulted on the Communication subdomain.

Content Validity

The next research question aimed at determining if the SSiS-RS item content represents social skills that are viewed by teachers as critical in the Australian samples in comparison to the US sample. To answer this question, the mean importance ratings of the Social Skills domain of the SSiS-RS were compared via independent samples t-tests at the domain and subdomain level between the US and the Australian samples (See Table 5). Importance ratings range from 0 to 2 for each item within the Social Skills domain. A rating of zero indicates that the teacher does not think a skill is important in the classroom, a rating of one indicates that the teacher believes the skill is important, and a rating of two indicates that the teacher believes the skill is critical in a classroom. Overall mean ratings for the Social Skills domain and subdomains ranged from 1.01 to 1.63 for the Australian sample and from 1.07 to 1.47 for the US sample.

Table 5

Comparison of Social Skills Importance Ratings Means of Australian and US Samples

SSiS Scale	Australian (n = 280)	United States (n = 940)		
	M (*SD*)	*M* (*SD*)	*p*	Hedges's *g*
Social Skills	1.22 (.24)	1.19 (.27)	.039*	.14
Communication	1.13 (.31)	1.13 (.31)	.947	
Cooperation	1.63 (.32)	1.47 (.36)	.000*	.47
Assertion	1.13 (.29)	1.07 (.30)	.003*	.20
Responsibility	1.33 (.32)	1.29 (.34)	.031*	.14
Empathy	1.01 (.31)	1.07 (.35)	.017*	.15
Engagement	1.13 (.27)	1.11 (.30)	.326	
Self-Control	1.22 (.32)	1.20 (.33)	.282	

Note. Levene's test was significant for the several of the domains (Cooperation, Responsibility, Empathy, and Engagement), suggesting that variances are not equal. In addition, samples sizes are not equal. Hedges's *g* was used to calculate effect size because equal variances were not assumed when calculating *t* for those domains.
$p < .05$

For the Social Skills domain area, the Australian teachers (M = 1.22, SD = .24) rated social skills as being more critical in their classrooms compared to US teachers (M = 1.19, SD = .27), t (1218) = 2.07, p =.039. The effect size, as measured by Hedges's g, was small (d = .14). In several subdomain areas (i.e., Cooperation, Assertion, Responsibility), Australian teachers rated the items as being more critical compared to the US teachers. Effect sizes ranged from .14 to .47. The largest effect size was found for the Cooperation subdomain, with the Australian mean importance rating equaling 1.63 (SD =.32) and the US mean equaling 1.47 (SD = .36). Although most of the significant

differences indicated that the Australian teachers rated social skills as being more important, this pattern differed for the Empathy subdomain. The US teachers (M = 1.07, SD = .35) rated these skills to be more important than the Australian teachers (M = 1.01, SD =.31), t (510.047) = -2.395, p =.017. The effect size for this difference was small (d = .15).

Internal Structure Validity: Confirmatory Factor Analyses

In order to provide evidence of internal structure validity, the factor structure of the SSiS-RS was measured for both the US and Australian samples using SPSS Amos. The proposed model was a three-factor (3F), first order model representing the three domains of the SSiS-RS. Given the poor results of this 3F model, an even more specified model, incorporating the subdomains of the SSiS-RS, was not attempted. In addition, a measurement invariance technique was not computed to test if there was a difference in the way the two sets of data fit the proposed model. Goodness of fit indices for the 3F model are summarized and compared in Table 6.

Table 6
Goodness of Fit Indices for 3F Model

	3F Australia (n = 280)	3F United States (n = 744)	Benchmark Values
NFI	.562	.640	>.95
CFI	.641	.683	>.90
GFI	.315	.455	>.90
RMSEA	.095	.080	<.08

Note. NF = Normed Fit Index; CFI = Comparative Fit Index; GFI = Goodness of Fit Index; RMSEA = Root Mean Square Error of Approximation; 3F = Three Factor Model

For the 3F model, the chi-square fit index was statistically significant with a value of 18914.24 (df = 3317, n = 744), $p < .05$ for the US data. Consistent with this result, fit indices also indicated less than adequate fit to the model, ranging from .46 to .68. That is,

the NFI (.64), the CFI (.68), and the GFI (.46) were all below the values proposed for adequate fit. The Standardized Root Mean Square Residual (Standardized RMSR) of .056 also indicated poor fit. However, the Root Mean Square Error of Approximation (RMSEA), at .08, is within the range that indicates good fit. The 3F model accounted for between 1% and 90% of the variance in each individual item. The saturated model had a lower AIC (6972.00) than did the 3F model (19252.239), indicating that the saturated model was a better fit, when not considering theory. The AIC of the Independence model (52753.19) was much higher than either.

Analyses from the Australian data produced similar results. The chi-square fit index was statistically significant with a value of 11643.776 (df = 3317, N = 280), $p <$ .05. Similarly, fit indices values also indicated less than adequate fit to the 3F model, ranging from .31 to .64. That is, the NFI (.56), the CFI (.64), and the GFI (.31) were all below the value proposed for adequate fit. The Standardized RMSR of .097 also indicated poor fit. In addition, the RMSEA of .095 is above the .08 value that would indicate good fit. The 3F model accounted for between .1% and 93% of the variance in each individual item. The saturated model had a lower AIC (6972.00) than did the three-factor model (11981.776), indicating that the saturated model was a better fit, when not considering theory. The AIC of the Independence model (26775.907) was much higher than either.

Factor loadings are important to consider as they can tell which items load best on each factor. Factor loadings below .5 may be inadequate. With regard to the Social Skills domain, the US data produced factors loadings ranging from .111 to .835 (see Table 7).

Table 7

Comparison of Factor Loadings for the Social Skills Domain

Item	Factor Loadings	
	3F US *n* = 744)	3F Aus (*n* = 280)
1. Ask for help.	.460	.255
2. Follows directions.	.758	.772
3. Comforts others.	.641	.673
4. Says "please."	.664	.746
5. Questions rules that may be unfair.	.111	.023
6. Is well-behaved when unsupervised.	.742	.744
7. Completes tasks without bothering others.	.724	.736
8. Forgives others.	.686	.723
9. Makes friends easily.	.455	.692
10. Responds well when others start a conversation or activity.	.723	.754
11. Stands up for herself/himself.	.194	.072
12. Participates appropriately in class.	.739	.809
13. Feels bad when others are sad.	.647	.711
14. Speaks in appropriate tone of voice.	.645	.678
15. Says when there is a problem.	.348	.232
16. Takes responsibility for his/her own actions.	.835	.786
17. Pays attention to your instructions.	.688	.791
18. Shows kindness to others when they are upset.	.760	.719
19. Interacts well with other children.	.743	.804
20. Takes turns in conversations.	.667	.800
21. Stays calm when teased.	.683	.722
22. Acts responsibly when with others.	.790	.794
23. Joins activities that have already started.	.451	.624
24. Says, "Thank you."	.686	.735
25. Expresses feelings when wronged.	.241	.118
26. Takes care when using other people's things.	.725	.789
27. Ignores classmates when they are distracting.	.661	.796
28. Is nice to others when they are feeling bad.	.710	.707
29. Invites others to join in activities.	.563	.659
30. Makes eye contact when talking.	.587	.648
31. Takes criticism without getting upset.	.611	.639
32. Respects the property of others.	.745	.774
33. Participates in games or group activities.	.491	.679

Table 7 continues

Table 7 continued

34. Uses appropriate language when upset.	.701	.758
35. Stands up for others who are treated unfairly.	.515	.514
36. Resolves disagreements with you calmly.	.645	.725
37. Follows classroom rules.	.745	.801
38. Shows concern for others.	.744	.706
39. Starts conversations with peers.	.412	.462
40. Uses gestures or body appropriately with others.	.686	.720
41. Responds appropriately when pushed or hit.	.744	.797
42. Takes responsibility for part of a group activity.	.755	.735
43. Introduces herself/himself to others.	.496	.496
44. Makes a compromise during a conflict.	.686	.791
45. Says nice things about herself/himself without bragging.	.517	.602
46. Stays clam when disagreeing with others.	.747	.730

The percentage of factor loadings below .50 for the US data is 20%. Factor loadings for the Australian data for the Social Skills domain ranged from .023 to .809. With regard to the Australian data, 15% of the factor loadings were below .50. Some low factor loadings for both samples include item 1 (asks for help), item 5 (questions rules that may be unfair), item 11 (stands up for him/herself), item 15 (says when there is a problem), item 25 (expresses feelings when wronged), and item 39 (starts conversations with peers). There were some factor loadings that were just below the .50 criteria for factor loadings in the US data but not in the Australian data, including participates in games or groups activities (item 33; .491), joins activities that have already started (item 23; .451), and makes friends easily (item 9; .455).

For the Problem Behaviors domain, factor loadings ranged from .264 to .795 (see Table 8) for the US sample. Approximately 17% of the factor loadings were below .50 in the US sample. For the Australian sample, factor loadings ranged from .242 to .831 for the Problem Behaviors domain. Approximately 27% of the factor loadings were below

.50 for the Australian data. Some of the low factor loadings in both samples included items 56 (withdrawing from others), 62 (getting embarrassed easily), 64 (acting lonely), 70 (having low energy), and 74 (acting sad or depressed). Other items that were below .50 in the Australian data included items representing following behaviors: stereotyped behaviors and acting anxious with others.

Table 8

Comparison of Factor Loadings for the Problem Behaviors Domain

Item	Factor Loadings	
	US (*n* = 744)	Aus (*n* = 280)
47. Acts without thinking.	.795	.718
48. Is preoccupied with object parts.	.595	.586
49. Bullies others.	.751	.797
50. Becomes upset when routines change.	.650	.617
51. Has difficulty waiting for turn.	.792	.825
52. Does things to make others feels scared.	.673	.757
53. Fidgets or moves around too much.	.717	.746
54. Has stereotyped motor behaviors.	.514	.375
55. Forces others to act against their will.	.690	.703
56. Withdraws from others.	.406	.458
57. Has temper tantrums.	.629	.800
58. Keeps others out of social circles.	.612	.683
59. Breaks into or stops group activities.	.779	.799
60. Repeats the same thing over and over.	.623	.518
61. Is aggressive toward people or objects.	.723	.815
62. Gets embarrassed easily.	.264	.246
63. Cheats in games or activities.	.742	.779
64. Acts lonely.	.423	.315
65. Is inattentive.	.696	.677
66. Has nonfunctional routines or rituals.	.624	.427
67. Fights well with others.	.762	.831
68. Says bad things about self.	.522	.582
69. Disobeys rules or requests.	.760	.813

Table 8 continues

Table 8 continued

70. Has low energy or is lethargic.	.338	.242
71. Gets distracted easily.	.712	.736
72. Uses odd physical gestures in interactions.	.592	.576
73. Talks back to adults.	.709	.767
74. Acts sad or depressed.	.492	.467
75. Lies or does not tell the truth.	.790	.780
76. Acts anxious with others.	.566	.436

Finally, the Academic Competence domain produced the highest factor loading values, ranging from .800 to .950 for the US sample. For the Australian data, the Academic Competence domain produced loading values ranging from .815 to .964. All factor loadings were much larger than the .50 value suggested for adequate factor loadings.

Table 9
Comparison of Factor Loadings for the Academic Competence Domain

Item	Factor Loadings	
	3F US (n = 744)	3F (n = 280)
77. Academic performance-compared to group.	.950	.964
78. Reading-compared to group.	.917	.920
79. Mathematics-compared to others.	.875	.943
80. Reading-Grade level expectations.	.914	.921
81. Mathematics-Grade level expectations.	.875	.938
82. Motivation to succeed.	.800	.815
83. Intellectual functioning-compared to others.	.899	.922

Overall, several of the factor loadings differed by more than .20 points between the two samples, indicating a potential difference in the way in which items load on the factors between the countries. All of the items that displayed this .20 difference were on the Social Skills domain and were below or near .50. The items that displayed this difference were the following: asking for help, expressing feelings when wronged, and making friends easily. For the first two items noted, the Australian sample produced

lower factor loadings compared to the US sample. For the last item listed, the Australian data produced higher factor loadings.

Mean Differences

PSGs means. The means for each of the PSGs categories were calculated for the Australian and US samples. Scores on the PSGs range from 1 to 5. While students who receive scores of 1, 2, or 3 are considered to be at-risk, students with a score of 4 or 5 are considered to not be at-risk. The mean scores ranged from 3.23 to 3.44 for the Australian data and 3.51 to 3.71 for the US data across the PSGs areas. As is depicted in Table 10, two of the PSG areas differed significantly between the two populations as indicated by independent sample *t* tests. In the Reading and Mathematics PSG domains, the US data resulted in higher mean values compared the Australian data. For the Reading PSG area, the mean difference between the United States (M = 3.71) and Australian (M = 3.26) samples resulted in a magnitude of effect in the medium range (g = .37), as measured by Hedges's g. As for the Mathematics PSG area, the magnitude of effect for the difference between the United States (M = 3.54) and Australian (M = 3.23) was in the small range (g = .29).

Table 10
Performance Screening Guides Means

Performance Scoring Guide Area	Mean (*SD*)		
	Australia (n = 857)	United States (n = 63)	p
Prosocial Behavior	3.40 (1.08)	3.51 (1.08)	.446
Motivation to Learn	3.44 (1.14)	3.67 (1.16)	.126
Reading	3.26 (1.23)	3.71 (1.20)	.004*
Mathematics	3.23 (1.09)	3.54 (1.06)	.030*

Note: Because Preschoolers are rated on a different scale, they were not included in this analysis.
*$p < .05$

SSiS-RS frequency means. Means were calculated for the SSiS-RS domains and subdomains for the South Brisbane and US samples. These means were compared via independent samples t-tests. Table 11 provides means and standard deviations for the domains and subdomains. The mean Social Skills domain standard score for the South Brisbane and US samples were 97.67 and 100.01, respectively. This difference between the two samples was not significant, t (1127) = -1.892, p = .080. The mean score for the Problem Behavior was 101.06 for the South Brisbane sample and 99.91 for the US sample. These mean scores did not differ significantly, t (1127) = .957, p = .352. In terms of the Academic Competence domain, the US produced a significantly higher mean frequency rating (M = 99.90) compared to the South Brisbane sample (M = 96.72). However, the magnitude of effect for this difference, as measured by Hedges's g was calculated to be in the small range (d = .21).

Differences between the samples in the subdomain areas were also assessed. It is important to note that unlike the domain areas, subdomain areas are not converted into standard scores. Within the domain of Social Skills, two subdomain areas differed between the South Brisbane and US samples. In the area of Engagement, the mean score for the South Brisbane sample (M =13.42) was lower than the mean score for the US sample (M = 14.63), t (1127) = -3.49, $p < .001$. The magnitude of effect, as measured by Hedges's g, was .31. Likewise, the Self-Control mean score was higher for the US sample (M = 14.13) compared to the South Brisbane sample (M = 13.23), t (1127) = -2.275, p = .024. The magnitude of effect for this difference was .19, which is in the small range. In the area of Empathy, the US sample (M = 12.29) produced a higher mean score than the South Brisbane data (M = 11.66), t (1127) = -2.035, p = .042. The magnitude of

effect was .17, which is in the small range. Within the Problem Behavior domain, several of the subdomains (i.e., Bullying, Hyperactivity, and Internalizing Behaviors) differed significantly between the two samples. In these cases, the US had lower mean values. The largest effect for these differences was large (.82) for the Bullying subdomain.

Table 11

SSiS-RS Frequency Ratings Means

SSiS Scale	*M* (*SD*)		
	South Brisbane (*n* =179)	United States (*n* = 950)	*p*
Social Skills	97.67 (16.59)	100.01 (14.91)	.080
Communication	15.51 (4.03)	15.79 (3.57)	.363
Cooperation	12.59 (4.06)	12.99 (3.57)	.222
Assertion	12.00 (3.87)	12.55 (4.21)	.104
Responsibility	12.83 (3.90)	13.31 (3.64)	.113
Empathy	11.66 (3.85)	12.29 (3.74)	.042*
Engagement	13.42 (4.32)	14.63 (3.80)	.001*
Self-Control	13.23 (4.91)	14.13 (4.47)	.024*
Problem Behaviors	101.06 (15.28)	99.91 (14.72)	.352
Externalizing	5.84 (6.46)	5.07 (6.15)	.128
Bullying	3.58 (4.42)	1.36 (2.27)	.000*
Hyperactivity/inattention	6.37 (4.94)	4.43 (4.39)	.000*
Internalizing	5.21 (3.84)	3.31 (3.35)	.000*
Autism Spectrum	9.88 (7.01)	8.61 (6.13)	.024*
Academic Competence	96.72 (15.5)	99.90 (15.03)	.012*

Note. For the Academic Competence domain, *n* = 744. Preschool children in this sample were not rated on this scale.
*p<.05

Internal Structure Validity: Correlations

To explore the internal structure validity of the SSiS, correlations between all domains on the PSGs and the SSiS-RS were calculated for the Australian and US samples using Pearson product-moment correlations. Table 12 displays the correlation matrices for the Australia and US samples for the PSGs and SSiS-RS domains.

Table 12

Pearson-product-moment correlations for Australian and United States Samples

Variable	PSG Pro Social	PSG MTL	PSG Reading	PSG Math	Social Skills	Problem Behavior	Academic Comp
1. PSG Pro Social	1	.74*	.44*	.49*	.67*	-.54*	.43*
2. PSG Motivation to Learn	.75*	1	.63*	.64*	.56*	-.53*	.55*
3. PSG Reading	.62*	.71*	1	.75*	.42*	-.24	.72*
4. PSG Math	.62*	.74*	.80*	1	.35	-.25	.66*
5. Social Skills	.77*	.71*	.56*	.60*	1	-.59*	.51*
6. Problem Behavior	-.67*	-.67*	-.50*	-.54*	-.73*	1	-.41*
7. Academic Comp	.65*	.73*	.83*	.82*	.60*	-.56*	1

Note. Values above the diagonal set of ones depict results from the United States data, while values below represent the results from the Australian data. For Australian sample, n = 280. For the US sample, correlations among PSGs and RS, n = 63 and correlations among RS, n =950. MTL= Motivation to Learn; PSG=Performance Screening Guides.
* $p < .05$, two-tailed

Relationships among PSG scores. Correlations among the PSG scores for both the US and Australian samples were all statistically significant, exhibiting large to very large positive relationships. For both samples, the Prosocial PSG scores demonstrated very large positive correlations with the Motivation to Learn PSG scores. In addition, the relationships between the Prosocial PSG scores and both the Mathematics PSG scores and the Reading PSG scores resulted in large correlations for the Australian sample and medium correlations in the US sample. The Motivation to Learn PSG scores demonstrated very large, positive relationships with the Mathematics and Reading PSG scores for the Australia sample, and large positive relationships for the US sample.

Finally, the correlations between the Mathematics and Reading PSG scores were very large and positive for both samples.

Relationships among SSiS-RS scores. Correlations among the SSiS-RS domain scores also exhibited expected results, demonstrating relationships within the medium and large ranges. For the Australian population, the Social Skills domain scores and the Problem Behavior domain scores revealed a very large, negative relationship, r = -.73. The relationship between the Academic Competence domain scores and the Problem Behavior and the Social Skills domain scores resulted in large, negative (r = -.56) and large, positive (r = .60) Pearson r values. The US correlations indicated similar relationships, albeit somewhat lower Pearson r values, among SSiS-RS domain scores.

Relationships between the PSG scores and the SSiS-RS domain scores. The relationship between the PSG scores and the SSiS-RS scores showed anticipated results. The Prosocial PSG scores correlated positively with the Social Skills and Academic Competence domains and negatively with the Problem Behavior domain for both the Australian and US samples. In terms of strength of the relationships, the Prosocial PSG scores resulted in very large and large correlations when correlated with the Social Skills domain scores, the Problem Behavior domain scores, and the Academic Competence domain scores, for the Australian sample. The absolute value of these correlations ranged from r = .65 to r = .77; the largest value was for the Social Skills domain. For the US sample, the correlations between the Prosocial PSG scores and each of the Rating Scales domain scores ranged from medium to large, with absolute values of the coefficients ranging from r = .43 to r =.67.

For both samples, the Motivation to Learn PSGs scores correlated positively with the Academic Competence domain and Social Skills domain scores and negatively with the Problem Behavior domain scores. The correlations between the Motivation to Learn PSG scores and the Academic Competence domain scores indicated very large relationships for the Australian data (r = .73) and a large relationship with the US data (r =.55). For the Australian sample, the correlation between the Motivation to Learn PSG scores and the Social Skills domain scores shared a very large relationship (r = .71), while the correlation for the US data indicated a large relationship (r =.56). The correlation between the Motivation to Learn PSG scores and Problem Behavior domain scores indicated large, negative relationships in both the Australian (r = -.67) and US (r = -.53) samples.

For the Australian sample, the correlations between the Mathematics/Reading PSG scores and the Problem Behavior and Social Skills domain scores demonstrated r values that were in the large range. The Mathematics and Reading PSG scores and the Academic Competence domain correlations were in the very large range, resulting in Pearson r values above .80 in the Australian Sample. For the US sample, the Reading/Mathematics PSGs and the Social Skills domain resulted in correlations in the medium range. When correlated with the Problem Behavior domain, Pearson r values were non significant for the US data. For relationships between the Reading/Mathematics PSGs and the Academic Competence domain, values were in the large and very large range for the US data.

Overall comparison. To further analyze the correlations, a multitrait-multimethod (Campbell & Fiske, 1959) organization of correlations was used to

determine if intercorrelations among the constructs measured on the different measures provided evidence of construct validity. Correlations were grouped based on similarity of construct, and based on similarity of a measure. For example, the first group included combinations of variables that measure similar constructs (e.g., Mathematics PSG and Academic Competence) but are from different measures (e.g., the PSGs and the SSiS-RS). The second group, which is hypothesized to have somewhat lower values, were combinations of variables that measured different constructs (e.g., Academic Competence and Socials Skills), but are from the same measure (e.g., the SSiS-RS). The last group should have the lowest values and are combinations that measure different constructs (e.g., Mathematics PSGs and Social Skills) and are from different measures. Table 13 provides a depiction of the different combinations for each group.

Table 13
Organization of Correlations by Groups

	Australian	US
Similar Construct-Different Measure		
PSGs Mathematics & AC	.82	.66
PSGs Reading & AC	.83	.72
PSGs Prosocial & SS RS	.77	.67
PSGs Motivation to Learn & SS RS	.71	.56
Different Construct-Same Measure		
PSGs Mathematics & PSGs Motivation to Learn	.74	.64
PSGs Mathematics & PSGs Prosocial	.62	.49
AC & SS	.60	.51
Reading PSGs and PSGs Motivation to Learn	.71	.63
PSGs Reading and PSGs Prosocial	.62	.44
Different Construct-Different Measure		
PSGs Prosocial & AC	.65	.43
PSGs Mathematics and SS	.60	.35
PSGs Reading and SS	.56	.42

Note. For Australian sample, n = 280. For the US sample, correlations between PSGs and RS scores, n = 63 and for correlations among SSiS-RS, n = 950. PSGs = Performance Screening Guides; SS = Social Skills; AC = Academic Competence.

The first set of correlations, including variables of similar constructs from different measures, resulted in Pearson *r* values in the very large range for the Australian data. Values ranged from *r* = .71 to *r* = .83. In the US data, correlations ranged from the large range to the very large range. Values ranged from *r* = .56 to *r* = .72. However, only one out of the four combinations had a value that was within the very large range. This correlation was the relationship between the PSGs Reading scores and the Academic Competence domain scores.

The correlations among variables of different constructs on the same measure resulted in two of the five correlations within the very large range for the Australian data. The rest were within the large range. With respect to the US correlations, all correlations were in the medium to the large range (*r* = .44 to *r* = .64).

This last group of correlations, including correlations among different constructs of different measures, should be lower than the previous two combinations. For the Australian sample, all three correlations were within the large range (*r* = .65*, r* = .60*,* and *r* = .56*)*. For the US correlations, all were in the medium range, ranging from *r* = .35 to *r* = .43.

Relationships among subdomains. Among the Social Skills subdomain areas, all correlations were significant and positive and ranged from small relationships to very large relationships in the US sample and from small to nearly perfect correlations in the Australian sample. The strongest correlations were between the Responsibility and Cooperation subdomains for both the Australian (*r* = .91) and US samples (*r* = .86). The weakest relationships were between the Assertion and Cooperation subdomains, resulting in Pearson *r* values in small range for the Australian (*r* = .23) and US samples (*r* = .30).

The relationships among the Problem Behavior subdomain areas resulted in significant, positive correlations. These correlations ranged from small to nearly perfection correlations. Externalizing and Hyperactivity areas showed the strongest Pearson *r* calculation (in the *r* = .90 range for both samples). This is not surprising given that several of the same items load on both subdomains.

Internal Structure Validity: Conditional Probability Analyses

The rationale for using a multiple gating assessment is that the more universal level of assessment will be more sensitive. Conditional probability analyses were conducted to provide information about the internal structure validity of the SSiS as a multiple gating assessment procedure in both the US and South Brisbane samples. The goal of this research question is to investigate how well children's scores on the PSGs predict children's scores on the SSiS-RS. To do this, conditional probability analyses were conducted for different combinations of PSGs and SSiS-RS scores. Table 14 provides a depiction of the combinations tested to provide evidence of internal structure validity.

Table 14
Depiction of Conditional Probability Analyses for Evidence of Internal Structure Validity

	Criterion (Rating Scales)			
	SS	PB	AC	Together
Predictor (PSGs)				
Prosocial	X	X		
Reading			X	
Math			X	
Motivation to Learn	X		X	
Together				X

Note. SS = Social Skills Domain; PB = Problem Behaviors Domain; AC = Academic Competence Domain; Together = falling at-risk on any of the areas

Results were similar between the two samples for all the indices (see Table 15). The Prosocial PSG scores were used to predict both the Social Skills and the Problem Behaviors domain scores. The Reading, Mathematics, and Motivation to Learn PSG scores were each used to predict scores on the Academic Competence domain.

Table 15

Conditional Probability Indices for the PSGs Predicting SSiS-RS

Screening System (Predictor → Criterion)	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
South Brisbane (*n* = 179)				
Prosocial PSG → Social Skills RS	.84	.68	.42	.94
Reading PSG → Academic Competence RS	.98	.69	.49	.99
Mathematics PSG → Academic Competence RS	.98	.68	.48	.99
Prosocial PSG → Problem Behavior RS	.89	.68	.40	.96
Motivation to Learn PSG → Academic Competence RS	.85	.71	.47	.94
Motivation to Learn PSG → Social Skills RS	.82	.69	.41	.93
PSG together → RS together	.92	.52	.53	.92
United States (*n* = 85)				
Prosocial PSG → Social Skills RS	.80	.62	.31	.94
Reading PSG → Academic Competence RS	.86	.80	.55	.95
Mathematics PSG → Academic Competence RS	.93	.70	.46	.97
Prosocial PSG → Problem Behavior RS	.87	.64	.34	.95
Motivation to Learn PSG → Academic Competence RS	.79	.65	.39	.91
Motivation to Learn PSG → Social Skills RS	.67	.67	.30	.90
PSG together →RS together	.96	.43	.51	.95

Note. PSG = Performance Screening Guide; RS = Rating Scales

For the Social Skills domain, Sensitivity was in the high range for both the US (.80) and South Brisbane (.84) samples. Thus, there is a high probability that the Prosocial PSG scores accurately identify a student with social difficulties as designated by the SSiS-RS for both samples. Specificity indices were in the moderate ranges for the US (.62) and South Brisbane (.68) samples. Positive Predictive Values were in the very low to the low ranges for the US (.31) and South Brisbane (.42) samples. Negative Predictive Values were in the high range for the US (.94) and South Brisbane (.94) samples. Note that a similar pattern resulted when the Motivation to Learn PSG scores were used to predict the Social Skills domain scores. Similar outcomes resulted when the Prosocial PSG was used as a predictor for the Problem Behavior domain scores. While Sensitivity (.87 and .89) and Negative Predictive Values (.95 and .96) were in the high range for both the US and South Brisbane samples, Specificity (.64 and .68) and Positive Predictive Values (.34 and .40) were in the very low to moderate ranges.

The Reading, Mathematics, and Motivation to Learn PSG scores were each used to predict scores on the Academic Competence domain. For the US sample, when the Reading PSG scores were used to predict the Academic Competence RS scores, Sensitivity (.86), Specificity (.80), and Negative Predictive Value (.95) indices were in the high range. The Positive Predictive Value was in the low range (.55). The South Brisbane data displayed a similar pattern of indices with Sensitivity (.98) and Negative Predictive Value (.99) indices in the high range, and Specificity (.69) and Positive Predictive Value (.49) indices in the moderate and low ranges, respectively. When the Mathematics PSG scores were used to predict the Academic Competence domain scores, analyses produced very similar results. For the US, Specificity (.93) and Negative

Predictive Value (.97) indices were in the high range, where as Specificity (.70) and Positive Predictive (.46) values were in the moderate and low ranges. The South Brisbane data produced the same pattern: Sensitivity (.98), Specificity (.68), Positive Predictive Value (.48), and Negative Predictive Value (.99). The Motivation to Learn PSG scores produced similar results when predicting the Academic Competence domain, with Sensitivity (.79 and .85) and Negative Predictive Value (.91 and .94) indices generally in the high range and Specificity (.65 and .71) and Positive Predictive Value (.39 and .47) indices generally in the low to moderate range for both US and South Brisbane samples.

When using the PSG scores together as the predictor (i.e., a student is considered to be at-risk if he/she was targeted on any of the PSGs) and the SSiS-RS scores together for the criterion variable (i.e., a student has a problem if he/she has been targeted on any of the SSiS-RS domains), Sensitivity indices were .92 and .96 for the South Brisbane and US populations, respectively. With respect to Specificity, values were .52 and .43, within the low range, for both the South Brisbane and US samples. Positive Predictive Values were in the low range for the South Brisbane data (.53) and the moderate range for the US data (.51). Negative Predictive Value indices fell within the high range for both the South Brisbane (.92) and US (.95) samples.

Relations to Other Variables

To provide evidence of predictive validity of the South Brisbane district sample, analyses to assess the relationships between the SSiS, particularly the PSGs, and the Second Year Diagnostic Net (hereafter Net) were conducted. The criterion variable, the Net, is a monitoring system in Queensland Australian based on teacher's judgments of

students' development in numeracy, reading, and writing. Correlations between the SSiS and the Net scores were first investigated for second graders in the South Brisbane school district (see Table 16).

Table 16
Person-product-moment correlations for South Brisbane Net Scores

Variable	Pro Social PSG	MTL PSG	Reading PSG	Mathematics PSG	Social Skills RS	Problem Behaviors RS	Academic Competence	Numeracy	Reading	Writing
Numeracy	.27*	.22*	.40*	.30*	.23	-.38*	.58*	1.00	-	-
Reading	.25*	.29*	.44*	.42*	.16	-.39*	.57*	.70*	1.00	-
Writing	.26*	.37*	.45*	.42*	.20	-.13	.42*	.50*	.58*	1.00

Note: MTL=Motivation to Learn; RS=Rating Scales. For this analysis, data was only available for second graders from the South Brisbane district (n =176). Available data Net data was available for n = 170. The number of second grad students who were also rated on the SSiS-RS was n = 60.
*p < .05, two-tailed

Relationships between Net scores and the Prosocial and Motivation to Learn PSG scores resulted in small to medium, positive correlations (ranging from r = .22 to r = .37) and medium sized positive correlations with the Mathematics and Reading PSG scores (ranging from r = .30 to r = .45). With regard to the relationships between the Social Skills domain scores and the Net scores, correlations indicated no significant relationships. The correlations between the Problem Behavior domain scores and the Net scores resulted in an insignificant relationship for the Writing Net scores and a medium, negative relationship with the Numeracy (r =-.38) and Reading (r = -.39) scores. Despite these small or non-significant results for the Social Skills and Problem Behaviors domains, the Academic Competence domain scores and the Numeracy (r = .58), Reading (r =.57), and Writing (r = .42) Net Scores resulted in medium to large correlations.

Conditional probability analyses. Conditional probability analyses were also conducted to provide evidence of predictive validity of the SSiS. The scores on the SSiS, particularly the PSGs, were used as the predictor variable and scores on the Second Year Diagnostic Net were used as the criterion variable (see Table 17).

Table 17
Conditional Probability Analyses among SSiS and Second Year Diagnostic Net

Screening System (Predictor → Criterion)	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
South Brisbane District				
Math PSG → Numeracy Net	.64	.70	.52	.79
Reading PSG → Reading Net	.54	.84	.78	.62
Reading PSG → Writing Net	.46	.84	.85	.45
Motivation to Learn PSG → Numeracy Net	.50	.72	.47	.74
Motivation to Learn PSG → Reading Net	.46	.76	.68	.56
Motivation to Learn → Writing Net	.44	.81	.81	.43
PSG Overall → Net overall	.69	.69	.86	.45
PSG + Academic Competence → Numeracy Net	.94	.39	.36	.94
PSG + Academic Competence → Reading Net	.89	.30	.59	.72
PSG + Academic Competence → Writing Net	.82	.41	.57	.72
PSG + Academic Competence → Overall Net	.79	.53	.80	.50

Note: For this analysis, data was only available for second graders from the South Brisbane district (n = 176). Available Net data was available for n = 170. The number of second grade students who were also rated on the SSiS-RS was n = 60.

Mathematics and Reading PSGs as predictor. The Mathematics PSG was used as a predictor for the Numeracy Net and the Reading PSG scores were used as the predictor for the Reading and Writing Net scores.

When the Mathematics PSG scores were used as the predictor for the Numeracy Net scores, indices ranged from the low to the moderate range. The Sensitivity (.64) and the Specificity (.70) indices were in the moderate range. The Positive Predictive Value was .52, which is in the low range. Lastly, the Negative Predictive Value was .79, which is in the moderate range.

The Reading PSG scores were used to predict both the Reading and the Writing Net scores. The Sensitivity indices were in the low range for predicting the Reading Net scores (.54) and for predicting the Writing Net scores (.46). Specificity indices were higher, demonstrating values in the high range (.84 for both the Reading and Writing Net). Positive Predictive values were in the moderate range for the Reading Net (.78) and the high range for the Writing Net (.85). Negative Predictive Value indices were in the low range for the Writing Net scores (.45) and moderate range for the Reading Net scores (.62).

Motivation to Learn PSG as predictor. When the Motivation to Learn PSG scores were used as predictors for each of the Net subject scores, Sensitivity indices ranged from .44 to .50, which are in the low range. In terms of Specificity, indices were moderate to high, ranging from .72 to .81. Positive Predictive Values differed depending on what criterion the Motivation to Learn scores were predicting. The values were .47, .68, and .81 for predicting the Numeracy, Reading, and Writing Net scores, respectively. Thus, while only 47% of the students identified as having a difficulty on the Motivation

to Learn PSG score did in fact show difficulty in reading, 81% of those students identified by the Motivation to Learn PSG had difficulty in writing. Negative Predictive Value indices were in the low to the moderate range when the Motivation to Learn PSG was used to predict the Numeracy (.74), Reading (.56), and Writing (.43) scores.

Overall PSGs and overall Net. When assessing how well being identified on any of the PSG sections predicted displaying difficulties on any of the Net areas, the index for Sensitivity and Specificity was in the moderate range (.69 for both). Positive Predictive Value was .86, indicating a high proportion of students who were targeted to be at-risk did in fact display academic difficulties. Negative Predictive Value was within the low range (.45).

The PSGs and SSiS-RS used together. When used together as a predictor, the PSGs and the Academic Competence Rating Scale score had higher indices for Sensitivity, resulting mostly in values in the high range. This indicates that when used together, the PSGs and Rating Scales are better at locating individuals with academic difficulties, as measured by the Net, compared to when the PSGs domains are used individually. However, these greater sensitivity scores seem to be at the expense of the Specificity scores, as they ranged from .30-.53, which are in the very low to low range. This indicates that using both the PSGs and the Academic Competence domain scores increases the ability to target students with difficulties, but lowers the ability to accurately not identify students without difficulties. Thus, it may over-identify students. In terms of Positive Predictive Value, probability indices ranged from the very low to the low range

for the Numeracy (.36), Reading (.59), and Writing (.57) Net areas. Negative Predictive Values ranged from the moderate to the high range for the Numeracy (.94), Reading (.72), and Writing (.72) areas.

Discussion

The purpose of this study was to examine the psychometric properties of the SSiS (Elliott and Gresham, 2008) in an Australian sample. This investigation was the second study (Kettler et al., 2011) designed to assess the psychometric properties of the SSiS in an Australian sample. Teachers from two different cities in Queensland, Australia rated their students on the PSGs and on the SSiS-RS. Several different types of psychometric information were examined from this data, including reliability, internal structure validity, content validity, and validity based on relations to other variables. When possible, comparisons to the US normative sample data were conducted. The SSiS is designed to be used as a multiple-gating assessment tool with a universal screening device, the PSGs, and a more targeted assessment measure, the SSiS-RS. The methods used to examine the psychometric qualities of the SSiS in both samples represent areas described in the Standards for Educational and Psychological testing (AERA, APA, & NCME, 1999).

Results indicated that reliability of the scores was moderately high to high for both samples across the domains and subdomains. For some of the domains, the reliability was statistically significantly higher in the Australian sample compared to the US sample. These differences were small, with a raw difference in reliability coefficients of less than .05. With respect to content validity, there were some differences with regard to what social skills each sample of teachers believed were more important. However,

effect sizes for these differences were generally small, indicating that for the most part, teachers from both samples assigned similar importance to social skills. With regard to internal structure validity, patterns of correlations among PSGs scores and SSiS-RS domain scores were similar between the two different countries. In addition, conditional probability analyses indicated that the PSGs accurately identify students with difficulties as measured by the SSiS-RS. The confirmatory factor analyses indicated that the three-factor model had poor fit to the data in both samples. There were few differences between the samples with respect to the factor loadings. Lastly, as concerns validity with respect to relations to other variables, the individual PSG scores demonstrated low to moderate ability to predict students who displayed difficulties on corresponding areas of the Net. Overall, the US and Australian data demonstrated similar results for the areas explored, consequently providing substantiation for the use of the tool in an Australian population.

Reliability

The first research question concerned how well the items of the SSiS fit together to generate the Social Skills, Problem Behaviors, and Academic Competence domain scores, and how well items fit together to generate subdomain scores for both samples. Results indicated that the reliability was moderately-high to high for both samples across subdomain and domain areas. In fact, all alpha coefficients exceeded .80 for both populations. In the literature it has been suggested that reliabilities above .80 are preferred for measures used in individual assessment (Sattler, 2008). The measure met this criterion for both samples. The values from the Australian data were larger than values found in a study that looked at the Social Skills Rating System in an Iranian

sample (Shahim, 2004). Shahim (2004) found alpha coefficients of **α** = .71 and **α** = .69 for the Social Skills and Problem Behavior domains, indicating only moderate reliability.

An independent samples Feldt test was used to identify significant differences in coefficient alpha values between the two samples. Results indicated that in many instances, the Australian sample produced larger alpha coefficient values. However, these differences were small, less than .05, suggesting that the differences may not be meaningful. One possible explanation for the small, albeit significant differences, may involve variability among scores in the Australian population. Levene's test was significant for several of the domain and subdomain scores, indicating that the populations differed with respect to variability. The Australian sample displayed higher variability among scores. According to Sattler (2008), greater variability in scores can contribute to larger reliability estimates. In summary, the internal consistency reliability of the SSiS-RS in an Australian sample was consistent with or exceeded that of the US standardization sample. Thus, these results provide evidence that the SSiS-RS is a reliable measure in this Australian sample.

The previous study utilizing the South Brisbane sample found similar results. Kettler et al. (2011) compared the South Brisbane reliability coefficients from third and fifth graders to the US standardization data. Results indicated that the reliability coefficients were similar between the two samples and were consistent with the results from the current study, which utilized a larger Australian sample. The current results expand on the results found in the 2011 study.

Mean Differences

Mean scores for the PSGs and the SSiS-RS were calculated and compared across samples. The PSG mean scores were calculated and compared to assess whether or not both samples rate their students equally on this universal measure. Results indicated that teachers rated students similarly across both samples for the Motivation to Learn and Prosocial PSG areas. For the Reading and Mathematics PSG areas, the United States data resulted in higher mean values. On the PSGs, a lower rating is indicative of more at-risk behavioral patterns in a particular PSG area. Thus, overall, the Australian teachers rated their students at being more at-risk on the academic areas of the PSGs.

Mean SSiS-RS scores were also calculated to investigate mean differences across samples. Results indicated that the Academic Competence domain was the only domain to show significant differences between the populations. Some of the subdomain scores resulted in significant differences between the samples. The differences on the Social Skills domain areas (e.g., Empathy, Engagement and Self-Control) resulted in higher values in the US data compared to the South Brisbane data. Following a similar pattern, in which the South Brisbane students were rated more negatively, the differences on the Problem Behaviors domain resulted in higher scores in the South Brisbane data for the Bullying, Hyperactivity/Inattention, Internalizing, and Autism Spectrum. The largest effect size was for the Bullying domain. This may suggest that overall the different samples vary in their experience of bullying in the classroom. It should be noted that the subdomain scores are raw scores. These scores are not converted into standard scores like the domain areas. Thus, age is not taken into account when calculating the frequency means for the subdomain scores. Whereas the South Brisbane sample contained students

only in the elementary school grades, the US sample had a wider sampling, containing students ranging from preschool to high school. These types of behaviors may be more common in the younger, elementary school grades.

Validity Inferences

Properties of the SSiS-RS were assessed to establish evidence of content validity and internal structure validity. Investigating the relationships between the PSGs and the SSiS-RS assisted in assessing the validity of the multi-phase system of the SSiS. Finally, investigating how accurately children's scores on the PSGs predicted scores on an academic monitoring system, the Year Two Diagnostic Net, assessed validity concerning relations to other variables.

Content validity. Investigating teachers' responses on the importance ratings helped to assess the appropriateness of the item content for the SSiS-RS Social Skills domain in the Australian sample. Results were compared across samples to determine whether there were any differences between the US and Australian samples. All ratings, across both samples, ranged from 1.01 to 1.63, indicating that mean importance ratings for all areas were between one and two. A rating of one for an item indicates that the teacher believes the skill is important, but neither critical nor non-important (scores of 0 and 2). The Australian sample had a statistically larger importance rating mean overall and for three of the subdomains. Overall, the Australian teachers rated items that aimed to measure cooperation, responsibility, and assertion as more important compared to the US sample. The largest effect size was for the items measuring cooperation. Interestingly, the US teachers rated items aimed to measure empathy as more important compared to the

teachers from the Australian sample. However, the effect size for this difference was small.

The most highly rated area appears to be the Cooperation subdomain, indicating that teachers in both samples may believe that the behaviors that relate to cooperation are some of the most important behaviors to display in the classroom. Gresham and Elliott (2008) found similar results: teachers' ratings indicated that three of the most important behaviors related to cooperation (i.e., listening to others, following directions, and following classroom rules).

The reason for the differences between the populations on the aforementioned areas warrants a discussion. Teachers from the Bundamba sample only rated children on the SSiS-RS who were targeted as being at-risk on the PSGs. On the SSiS-RS, importance ratings are linked to ratings of individual students in that for every social skills item, a student gets a frequency rating and an importance rating. Thus, teachers rate multiple students and provide an importance rating for each student for every item. One may reason that the teacher may have been associating the skills with a particular student who actually received a more problematic rating scale score compared to the general population. However, further analyses indicate that the Australian samples did not differ in their importance ratings overall, suggesting that teachers in the Bundamba sample did not differ in their importance ratings despite rating only the more at-risk students.

Internal structure validity. To assess the internal structure validity of the SSiS in the Australian sample, several analyses were conducted, including a confirmatory factor analysis, correlational analyses, and conditional probability analyses. Results from the analyses were similar for the US and Australian samples.

Relations among SSiS-RS items. The goal of this analysis was to test how well the data from both samples fit the factor model, as depicted in the scoring manual and structure of the SSiS-RS. The structure of the SSiS-RS is a three-factor model including a Social Skills domain, a Problem Behaviors domain, and an Academic Competence domain. The three-factor model exhibited poor fit for both samples. Given that this basic model did not fit the data well, no additional models (i.e., a second order model including the subdomains) were tested. In addition, a measurement invariance technique was deemed unnecessary given that the first step in such analyses is to identify baseline models that fit the data well (Brown, 2006).

Looking at the individual domains, the best fitting part of the model appears to be the Academic Competence domain with factor loadings ranging from .82 to .96. This is not unexpected given that these items are all very similar, relating to somewhat objective criteria (e.g., ranking students based on their academic achievement). There were no substantial differences in the factor loadings for any of the items on this domain between the two samples. These items may be easier to interpret and thus rate.

After investigating the low factor loadings on the Problem Behavior scale, it appears that the items with the lowest factor loadings concern internalizing or withdrawing behaviors. Conversely, all other items seem to concern more disruptive or distracting behavior patterns. The combination of the different, and sometimes opposing, problem behaviors may contribute to these low factor loadings. On the Social Skills domain, many of the low factor loadings were items that fall on the Assertion subdomain, indicating that scores on these items may not fit with scores on other items that fall on the social skills domain. The only substantial differences in factor loadings between the two

samples were items on the Social Skills domain. For example, for the US data, the item, "Asks for help" led to a higher, albeit still low, factor loading compared to the Australian data. On the contrary, the item, "Makes friends easily" had a higher factor loading in the Australian data compared to the US data. Finally, the item "Expressing feelings when wronged" was below the .5 criteria for both samples, but the US data produced a slightly higher factor loading.

Other studies that assessed the internal structure validity of different rating scales via confirmatory factor analyses found more positive results (e.g., Danielson & Phelps, 2003). There are several possible reasons for the weak model fit to the data in both the US and Australian samples. The SSiS-RS is a direct revision of the Social Skills Rating Scales (Elliott and Gresham, 1990). New items were added to each of the domains. The creators of the SSiS note in the manual that confirmatory factor analysis was used to create beneficial changes to the Social Skills scale. However, when given the results of the confirmatory factor analysis, the authors decided not to make all the beneficial changes in order to preserve the goal of the revisions (i.e., higher reliability). Although the creators of the tool used confirmatory factor analysis to construct the Social Skills domain, it was not tested when conducting their validity studies and not used as a criterion to substantiate the validity of the tool. There was no information concerning whether confirmatory factor analysis was used on the other domain areas.

Relations among SSiS-RS domain areas and PSGs. Although the evidence of internal structure validity from the confirmatory factor analyses was not strong for either sample, other evidence of internal structure validity had more positive results. Magnitude and direction of correlations were similar across samples. All of the relationships among

the PSGs and the SSiS-RS domain areas resulted in significant correlations, indicating that all areas of the SSiS are significantly related. The pattern of correlations indicated that the constructs are being measured appropriately, according to a multitrait-multimethod framework (Campbell & Fiske, 1959). For both samples, there were slightly stronger correlations among scores measuring the same construct on different measures (e.g., the Prosocial Behavior PSG and the Social Skills domain) compared to correlations measuring different constructs on the same measures (e.g., Social Skills domain and Academic Competence).

The final way in which internal structure validity evidence was collected was through conditional probability analyses. For these analyses, the PSGs were used as predictors for the SSiS-RS. Sensitivity indices ranged from .82 to .98 and .79 to 1 in the South Brisbane and US samples, respectively. These results indicate that Sensitivity was within the high range for most of the combinations tested. Thus, if a student has academic, social, and/or behavior problems as designated by the SSiS-RS domains, then there is a high probability that the PSGs will identify the student as having these problems. Specificity values, however, displayed a different pattern of results. Specificity indices ranged from the low range to the moderate range for nearly all combinations tested for both samples. Based on these results it appears as though PSGs were less likely to correctly not identify students without a difficulty. Positive Predictive Values were in the low range for all the combinations for the South Brisbane sample. The US data resulted in a similar pattern. This indicates that the likelihood that a student who was identified on the PSGs did in fact display difficulties as measured by the SSiS-RS was low. Negative Predictive Value indices resulted in values in the high range for all

combinations in both the South Brisbane and US samples. This indicates that there was a high likelihood that a student who was not identified on the PSGs was a student without a difficulty as measured by the SSiS-RS.

These results indicate that the PSGs work appropriately as a first step in the multi-stage rating system as intended for both samples. More specifically, the Prosocial PSG appropriately identifies students displaying difficulties on the Problem Behavior and Social Skills domain. The Math/Reading PSG appropriately identifies students who display difficulties on the Academic Competence domain. The results suggest that for both samples, these PSG areas can be used to screen a classroom and identify students who may need more in-depth assessment in these areas. Positive Predictive Value and Specificity values fall below the high range, this may be acceptable for this type of screening system, in which a false positive is not very costly (Kettler & Feeney-Kettler, 2011). If a student were inaccurately identified as at-risk on the PSGs, more targeted analyses would be conducted to determine the nature of the difficulty. On the more extensive measure, the SSiS-RS, the student may not be found to display difficulties and no further intervention would be utilized.

In fact, with a universal screening system as a first step of a multiple-gating assessment procedure, it is best to over-identify, rather than under-identify students who display difficulties. On the one hand, over-identification in screening systems can lead to certain stressors related to serving misidentified students, including overuse of resources for programming and stress among family and/or support personnel (Glover and Albers, 2007). On the other hand, students who are under-identified may not receive the necessary supports, services, and more intensive assessment measures. Different stages of

a multiple-gating assessment procedure require different expectations with regard to identification. Glovers and Albers (2007) indicate that sensitivity (or inclusion) as a first step in a multiple-gating assessment procedure is critical. In fact it is noted that, "it may be useful to compromise precision (e.g., a high positive predictive value or specificity) for inclusion (sensitivity)" (Glovers and Albers, 2007). At later stages of a multiple-gating procedure, higher precision is expected.

Relations to other variables. To explore how the SSiS relates to other variables, correlations were computed to assess the relationships between the Year Two Diagnostic Net and the SSiS areas. The Year Two Diagnostic Net is a monitoring system used in Queensland, Australia. Students' ratings are based on teachers' judgments of students' development in numeracy, reading, and writing. Conditional probability analyses were conducted to assess how well the PSGs and Academic Competence Rating Scale scores predicted achievement on the Net for second graders in the South Brisbane district. Correlations among Net scores and PSG scores indicate trivially small to large relationships, depending on the area being correlated. For the areas that are less associated with academics (i.e., Prosocial and Motivation to Learn), relationships with the Net correlations were generally small. For the more direct academic PSG areas (i.e., Mathematics and Reading), correlations were in the medium range. Likewise, the Academic Competence domain displayed medium to large correlations with the Net scores, while the Social Skills and Problem Behaviors domain were shown to have non-significant, small or medium correlations. As expected, it appears as though the academic areas of the SSiS-RS related better to the Net scores than did the areas related to social skills or problem behaviors.

Conditional probability analyses were used to determine how well at-risk ratings on the PSGs predicted the presence of an academic difficulty as identified by the Year Two Diagnostic Net. Results indicated that the Math PSGs showed moderate accuracy in identifying students who actually displayed academic difficulties, indicated by a Sensitivity value of .64. All other PSGs had low accuracy in identifying students who actually displayed academic difficulties on the Net. Specificity values were all within the moderate range when the individual PSGs were used to predict corresponding areas on the Net. This trend indicates that the PSGs seemed to have more accuracy in correctly not identifying students without difficulties compared to correctly identifying students who display academic difficulties. Positive Predictive Values ranged from the low range to the high range. The highest Positive Predictive Value was for the Reading PSG scores predicting the Writing Net scores. The lowest values resulted when the Numeracy Net was the criterion variable. Negative Predictive Values ranged from the low range to the moderate range.

When the PSG was used together with the Academic Competence domain of the SSiS-RS, results indicated that the combination tended to over-identify students with difficulties. While Sensitivity indices were in the high range, Specificity indices were in the very low to the low range. Thus, although the PSGs and the Academic Competence domain were able to correctly identify a large proportion of students who displayed difficulties in these academic subjects, there was a low likelihood that the SSiS correctly did not identify students who did not display difficulties. Because the SSiS is meant to be a system to identify students at-risk, the false positives are expected and are not as costly as false negatives. However, using both the PSGs and the Academic Competence domain

often resulted in over half of the students being identified as at-risk, which contributed to the pattern of results mentioned above.

Kettler et al. (2011) investigated the predictive validity of the SSiS-RS in the South Brisbane data using conditional probability analyses and multiple regression analyses. It was concluded that the PSGs and the Rating Scales are both predictive of academic achievment and work best when used together in this Australian sample. In the current study, the measure of academic achievement was the Net, a teacher judgement based measure. The current study found that the PSGs and the Rating Scales together better predicted the presence of an academic difficulty, resulting in higher Sensitivty values. However, the current study indicates that the combination may over-identify students, resulting in lower Specificity values.

Limitations

There are some limitations in the methods of this study. First, the Australian dataset was the product of previously collected data from two separate studies. At times, this limited the analyses that could be computed and complicated the procedures involved in the analyses. Because the students from the Bundamba subsample were only rated on the SSiS-RS if they were at-risk on the PSGs, this subsample was removed from several of the analyses (e.g., mean differences, conditional probability analyses). In addition, because teachers rated multiple students on the SSiS, nesting and dependency among scores provided by the same teacher may also be a limitation in this study. In the Australian data, teachers rated an entire classroom of students on the SSiS-RS. This nesting may violate the assumption of the independence among rating scores on the measures.

Moreover, the way in which the data were collected could not be equivalent to that of the US standardization data. This limits the conclusions that can be drawn from the differences between the two samples. Results cannot be generalized to the entire Australian population, since Queensland may not be representative of Australia as a whole. Unlike the US data, which spanned different regions of the country, the Australian data only spanned one region, thus limiting the generalizations that can be inferred from the results. In addition, the Australian sample only spanned elementary school grades. The results provide evidence for this sample but do not generalize to an entire population. Before suggesting that such a tool can be adequately used in an entire country, research will need to target a more representative sample in terms of student age and geographical region.

The evidence based on relations to other variables may also be limited in the information that it provided given that the Year Two Diagnostic Net is no longer used. Because of this, data could not be collected for the Bundamba data and thus the sample size for this analysis was small. In addition, there is not much information regarding the reliability and the validity of the Net itself, which limits the findings of this portion of the study. Furthermore, the Net provides information on another measure of teacher judgment, much like the SSiS, rather than a standardized achievement measure. Using a more reliable measure from which valid inferences can be drawn about achievement and comparing the results across the samples could extend this research.

Implications for Future Research and Practice

The results of this study indicate that the SSiS is an adequate tool for use in this Australian sample. Most of the evidence suggests that the psychometric properties are

similar to that of the US sample. In addition, there is evidence that the tool works as a multi-tiered assessment for identifying students that display academic, behavioral, and social skill difficulties in the classroom.

Future studies should investigate the psychometric properties of the SSiS in other samples and countries. In addition, this study did not assess the usability of the tool for the Australian population. Usability is another area to explore when evaluating the psychometric properties of a measure in a sample (Kettler & Feeney-Kettler, 2011). In the literature on assessment tools, usability has been defined as consisting of the following characteristics: (a) cost of administration; (b) feasibility, including clarity of instructions; (c) acceptability to the multiple stakeholders; (d) required infrastructure for managing the data; (e) availability of appropriate accommodations; and (f) connection to improved treatment utility (Glover and Albers, 2007). A future study could assess perceptions of teachers, practitioners, and other stakeholders when using the SSiS-RS and PSGs. Input and buy-in are important when considering the adoption of a program or a tool in a system (Rogers 2003). Future research might compare these perceptions across US and Australian samples.

Moreover, it may be beneficial for future research to investigate the effects of using this tool to develop and implement interventions. The Social Skills Intervention Guide, a feature of the SSiS, is a resource for teachers and practitioners to provide social skills training to students based on results from the assessment. The implementation and intervention efficacy of this tool should be a topic of future research. The SSiS-RS as a pre/post measure to assess intervention effects in the classroom may also be a beneficial

focus for future research. These are important and innovative features of the SSiS that are not addressed in the current study.

Conclusion

In summary, the results demonstrated that the reliability and validity evidence assessed in this study is comparable between the US and Australian samples for the SSiS-RS and PSGs. The measure demonstrates moderately high to high reliability for the rating scales in both samples. Evidence for the internal structure had mixed results in that the confirmatory factor analysis indicated that neither set of data fit a 3-factor model. However, other evidence of internal structure validity, evidenced by correlations and conditional probability analyses, was more promising. The high reliability results are commensurate with the goals of the revision of the Social Skills Rating system, whereas a good factor structure was not an outlined goal. Overall, these findings indicate that the SSiS is an appropriate tool to use in Australian samples of elementary school students.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Australian Curriculum, Assessment and Reporting Authority. (2011). *General Capabilites: ACARA*. Retrieved March 5, 2013, from Home: ACARA: http://www.acara.edu.au/home_page.html

Bentler, P.M (1990). Comparative fit indices in structural models. *Psychological Bulletin*, 107, 238-246.

Campbell, Donald T. (1959). Convergent and discriminant validation by the multitrait multimethod matrix, *Psychological Bulletin, 56*, 81-105.

Cheong, Y.F., & Raudenbush, S.W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods,* 5(4), 477-495.

Cortina, J. M. (1993). What is alpha? An explanation of theory and applications. *Journal of Applied Psychology , 78* (1), 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Danielson, C. K., & Phelps, C. R. (2003). The assessment of children's social skills through self- report: A potential screening instrument for classroom use. *Measurement and Evaluation in Counseling and Development*, 35, 218-229.

DiPerna, J.C., & Elliott, S.N. (1999). The development and validation of the Academic Competence Evaluation Scale. *Journal of Psychoeducational Assessment*, *17*, 207-225.

DuPaul, G.J., Anastopoulos, A.D., Power, T.J., Reid, R., Ikeda, M.J., & McGoey, K.E. (1998). Parent ratings of Attention-Deficit/Hyperactivity Disorder symptoms: Factor structure and normative data. Journal of Psychopathology and Behavioral Assessment, 20, 83-102.

Gencdogan, B. (2008). Psychometric properties of the Turkish version of the childrens' self-report social skills scale. *Social Behavior and Personality,* 36 (7), 955-964.

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire (SDQ). Journal of the American Academy of Child and Adolescent Psychiatry, 40, 1337-1345.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, *45*(2), 117-135.

Gresham, F., & Elliott, S. (2008). *SSiS: Teacher Rating Scales.* Bloomington, MN: Pearson Assessments.

Gresham, F.M., Elliott, S.N., Vance, M.J. & Cook, C.R. (2011). Comparability of Social Skills Rating System to the Social Skills Improvement System: Content and psychometric comparisons across elementary and secondary age levels. *School Psychology Quarterly, 26*, 27-44.

Gresham, F., & Elliott, S. (1987). The relationship between adaptive behavior and social skills: Issues in definition and assessment. *The Journal of Special Education , 21* (1), 167-181.

Gresham, F., & Elliott, S. (1990). Social Skills Rating System manual. Circle Pines, MN: AGS.

Gresham, F., & Noell, G. (1996). Teachers as judges of social competence: A conditional probability analysis. *School Psychology Review , 25* (1), 108-118.

Hawes, D.J. & Dadds, M.R. (2004). Australian data and psychometric properties of the strengths and difficulties questionnaire. *Australian and New Zealand Journal of Psychiatry, 38* (8), 644-651.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. Review of Educational Research, 59, 297—313.

Ikeda, M.J., Neessen, E., & Witt, J.C. (2008). Best practices in universal screening. In A. Thomas & J. Grimes (Eds.), Best practices in school psychology V (pp.103-114). Bethesda, MD: National Association of School Psychologists.

Joreskog, K.G. & Sorbom, D. (1986). Lisrel VI. 4th Ed. Mooresville, Indiana: Scientific Software, Inc.

Kettler, R. J., & Feeney-Kettler, K. A. (2011). Screening systems and decision making at the preschool level: Application of a comprehensive validity framework. *Psychology in the Schools , 48* (5), 430-441.

Kettler, R. J., Elliott, S. N., Davies, M., & Griffin, P. (2011). Testing a multi-stage screening system: Predicting performance on Australia's national achievement test

using teachers' ratings of academic and social behaviors. *School Psychology International , 33* (1), 93-111.

Lane, K., Givner, C., & Pierson, M. (2004). Teacher expectations of student behavior: Social skills necessary for success in elementary school classrooms. *The Journal of Special Education , 38* (2), 104-110.

Liu, Jianghon (2004). Childhood externalizing behavior: Theory and implications. *Journal of Child and Adolescent Psychiatric Nursing, 17 (3), 93-103.*

Loehlin, J.C. (2004). Latent Variable Models: An Introduction to Factor, Path and Structural Equation Analysis. Lawrence Erlbaum Assoc Inc.

Lyon, M. A., Albertus, C., Birkinbine, J., & Naibi, J. (1996). A validity study of the social skills rating system-teacher version with disabled and nondisabled preschool children. *Perceptual and Motor Skills , 83*, 307-316.

Malecki, C. K., & Elliott, S. N. (2002). Children's social behaviors as predictors of academic achievment: A longitudunal analysis. *School Psychology Quarterly , 17* (1), 1-23.

Matson, J.L. & Wilkins, J. (2009). Psychometric testing methods for children's social skills. *Research in Developmental Disabilities, 30,* 249-274.

Meier, C. R., DiPerna, J. C., & Oster, M. M. (2006). Importance of social skills in the elementary grades. *Education and Treatment of Children , 29* (3), 409-419.

Merrell, K. W. (2001). Assessment of children's social skills: Recent developments, best practices, and new directions. *Exceptionality , 9* (1&2), 3-18.

Meyers, L., Gamst, G., & Guarino, A. (2006). *Applied Multivariate Research: Design and Interpretation.* London: SAGE Publications.

Pedersen, J. A., Worrell, F. C., & French, J. L. (2001). Reliability of the social skills rating system with rural appalachian children from families with low incomes. *Journal of Psychoeducational Assessment , 19*, 45-55.

Queensland Government: Department of Education, Training and Employment. (1998). *Year Two Diagnostic Net*. Retrieved January 1, 2013, from Education Queensland: http://education.qld.gov.au/

Rivera, B. D., & Rogers-Adkinson, D. (1997). Culturally sensitive interventions: Social skills traning with children and parents from culturally and linguistically diverse backgrounds. *Intervention in School and Clinic , 33* (2), 75-80.

Rydell, A.-M., Hagekull, B., & Bohlin, G. (1997). Measurement of two social competence aspects in middle childhood. *Developmental Psychology , 33* (5), 824-833.

Shahim, S. (2004). Reliability of the social skills rating system for preschool children in Iran. *Psychological Reports , 95*, 1264-1266.

Sparrow, S. S., Cicchetti, D., & Balla, D. A. (2005). *Vineland Adaptive Behavior Scales - 2nd Edition manual.* Minneapolis, MN: NCS Pearson, Inc.

Teo, A., Carlson, C., Mathieu, P., Egeland, B., & Sroufe, L. A. (1996). A prospective longitudinal study of psychosocial predictors of achievement. *Journal of School Psychology,* 34, 285-306.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement , 56* (2), 197-208.

Walthall, J. C., Konold, T. R., & Pianta, R. C. (2005). Factor structure of the social skills rating system across child gender and ethnicity. *Journal of psychoeducational assessment , 23*, 201-215.

Welsh, M., Parke, R., Widaman, K., & O'Neil. (2001). Linkages between children's social and academic competence: A longitudinal analysis. *Journal of School Psychology, 39* (6), 463-481.

Wentzel, K. (1993). Does being good make the grade? Social behavior and academic competence in middle school. *Journal of Educational Psychology , 85* (2), 357-364.

Zhang, D., Faries, D.E., Vowles, M., & Michelson, D. (2005). ADHD Rating Scale IV: Psychometric properties from a multi-national study as a clinician-administered instrument. *International Journal of Methods Psychiatric Research, 14,* 186-201