Global Journals

Global Journal of HUMAN-SOCIAL SCIENCE: G
Linguistics & Education
Volume 15 Issue 1 Version 1.0 Year 2015
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-460x & Print ISSN: 0975-587X

# The Impact of Language Translation on the Internal Structure of a Rating Scale: The Strengths and Difficulties Questionnaire in Spanish

By Heather Blumert, Ryan J. Kettler & Kimberley D. Lakes

*Rutgers, The State University of New Jersey, United States*

*Abstract-* The purpose of this study was to compare the psychometric properties of the Spanish version of the Strengths and Difficulties Questionnaire (SDQ), a 25-item behavioral screener, with the English version. Participants included in this study were 363 English-speaking parents and 334 Spanish-speaking parents of preschool age children (ages 3-5) who took part in the Children's Hospital of Orange County/University of California (Irvine) Initiative for the Development of Attention and Readiness (CUIDAR) program from 2004-2008. This study used data from the CUIDAR program to explore mean rating differences between the English and Spanish versions of the SDQ, along with coefficient alpha as an indicator of reliability at the scale and composite level, and factor analytic evidence of score validity. Mean ratings of the scales and the Total Difficulties scale were very similar across language forms. Reliability coefficients indicated alphas were higher for scores derived from the English forms compared to the Spanish forms at the scale and composite levels, although neither form produced scores with adequate reliability at the scale level. Finally, the Five First Order Factor Model was the best-fitting and most valid representation of all 25 items of the SDQ, regardless of the language of the form.

*GJHSS-G Classification : FOR Code: 200499*

# The Impact of Language Translation on the Internal Structure of a Rating Scale: The Strengths and Difficulties Questionnaire in Spanish

## Internal Structure of the SDQ-Spanish

Heather Blumert [α], Ryan J. Kettler [σ] & Kimberley D. Lakes [ρ]

*Abstract-* The purpose of this study was to compare the psychometric properties of the Spanish version of the Strengths and Difficulties Questionnaire (SDQ), a 25-item behavioral screener, with the English version. Participants included in this study were 363 English-speaking parents and 334 Spanish-speaking parents of preschool age children (ages 3-5) who took part in the Children's Hospital of Orange County/University of California (Irvine) Initiative for the Development of Attention and Readiness (CUIDAR) program from 2004-2008. This study used data from the CUIDAR program to explore mean rating differences between the English and Spanish versions of the SDQ, along with coefficient alpha as an indicator of reliability at the scale and composite level, and factor analytic evidence of score validity. Mean ratings of the scales and the Total Difficulties scale were very similar across language forms. Reliability coefficients indicated alphas were higher for scores derived from the English forms compared to the Spanish forms at the scale and composite levels, although neither form produced scores with adequate reliability at the scale level. Finally, the Five First Order Factor Model was the best-fitting and most valid representation of all 25 items of the SDQ, regardless of the language of the form.

## I. Introduction

In the United States, Latinos represent the largest ethnic minority group (Pedrotti & Edwards, 2010), are overrepresented in terms of families afflicted by behavioral disorders and mental health disorders (Smokowski, Reynolds, & Bezruczko, 1999), and are at greater risk of failing in school as well as dropping out of school (Tinkler, 2002). Researchers (e.g., Lakes, Lopez, & Garro, 2006) have noted that to address such mental health disparities, it is important to develop and study clinical assessment methods in the populations in which they will be used. Recent research (Lakes, in press) illustrated how sample characteristics impact the reliability of scores obtained, providing further evidence of the importance of carefully studying assessment instruments in different populations before applying them widely or assuming that the psychometric properties of scores derived from these instruments will be equivalent in different populations.

As the Latino population and the number of Latino school-age children increase in numbers throughout the United States, it is essential to have instruments for Spanish-speaking individuals that will provide reliable and valid assessments of children's behavioral strengths and weaknesses. It is particularly important to understand the Latino parent perspective when they are asked to rate their children's behaviors. For many of these parents, Spanish is the only language in which they are fluent. Thus, there is a need for a measure in Spanish that identifies children's behavioral strengths and difficulties as well as the English version works for English-speaking families. The current study examines the psychometric properties of scores derived from a behavioral screening measure (Strengths and Difficulties Questionnaire, Goodman, 2001) that was first written in English, but has been translated to Spanish and is now widely used in both languages.

## II. Criteria for Evaluating Rating Forms

Exploring the psychometric properties of scores obtained from rating scales that have been translated into Spanish is essential. Key aspects in exploring the psychometric properties of a test or scale entail evaluating how reliable and valid its scores are.

Reliability refers to the how consistent a measure is when the assessment is repeated on a population (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999), and establishing reliability evidence is a prerequisite to establishing evidence for the validity of inferences drawn from scores. Coefficient alpha is one indicator of reliability, equal to the mean of all split-half reliabilities, when the standard deviations are equal (Cortina, 1993).

Validity refers to the degree to which theory and evidence provide backing for the interpretations of test

*Author α σ : Rutgers, The State University of New Jersey.*
*e-mail: r.j.kettler@rutgers.edu.*
*Author ρ: University of California – Irvine.*

scores entailed by the designed use of tests (AERA at al., 1999). Factor analysis is often used to provide evidence of how well the items on a scale fit together as intended, yielding one type of evidence for validity that is included in the *Standards for Educational and Psychological Testing* (the *Standards;* AERA et al., 1999). Exploratory factor analysis (EFA) is appropriate when no model is hypothesized before analysis, but when a model is theorized, confirmatory factor analysis (CFA) is a stronger evaluative tool. In CFA, the fit of each proposed model is tested to determine the best structure of a test (Sharkey et al., 2009). Subsequent links between validity and factor analysis lie in the theory of falsification, which posits that that a theory should not be considered credible until efforts have been made to disconfirm the theory (Thompson & Daniel, 1996). A strong program of construct validation requires that rival hypotheses be tested which may suggest alternative explanations for the meanings of test scores. Similarly, in CFA, rival models can and should be tested because multiple models may fit the same data. Multiple models are evaluated in the current study.

## III. Psychometric Properties of Assessment Tools Translated into Spanish

Research regarding the effect of translating instruments into Spanish, or other languages, has yielded varying results. The effect of translation differs by measure.

*The Behavioral and Emotional Rating Scale-2.*

The Behavioral and Emotional Rating Scale-2 Parent Report (BERS-2) is a school-based scale that measures the strengths of a student (Sharkey et al., 2009). It is used primarily with children who have significant mental health concerns, including Attention Deficit Hyperactivity Disorder (ADHD), Oppositional Defiant Disorder (ODD), and mood disorders. Buckley, Ryser, Reid, and Epstein (2006) performed an exploratory factor analysis of the original English version of the BERS-2. They assessed various factor structures, including a 3-factor model and the intended 5-factor structure, finding the 5- factor structure to be the best-fitting model (Buckley, 2006). Sharkey et al. (2009) then explored the factor structure of the BERS-2 with Spanish-Speaking parents of at-risk youth. There were two samples included in this study. The first consisted of parents of students in fourth through seventh grade from low socioeconomic status neighborhoods in two school districts in Central California. The second sample consisted of parents of youths enrolled in a community program providing services to criminally involved families. Exploratory factor analysis indicated that a three-factor model was a better fit than the original five-factor model of the English version for the latter sample.

*The Social Anxiety Scale for Adolescents (SAS-A).*

The Social Anxiety Scale for Adolescents (SAS-A) is an instrument designed to measure social anxiety responses (Olivares, Ruiz, Hidalgo, Garcia-Lopez, Rosa, & Piqueras, 2004). CFA of the SAS-A by LaGreca and Lopez (as cited in Olivares et al., 2004) supported the original three-factor structure in an English-speaking sample. Olivares et al., (2009) assessed alternative models to the original three-factor model of the SAS-A: a null or independent model, a one-factor model in which all 18 items loaded onto a general social anxiety factor, a two-factor model of FNE and SAD combined, and the original model, with a Spanish-speaking adolescent population in Spain. Results indicated that the three-factor model was confirmed and was a better fit compared to the alternative structures proposed. The three-factor model had the highest Goodness of Fit Index (.89) and Comparative Fit Index (.89) among the tested models. In addition, the Standardized Root Mean Square Residual of .06 indicated a good fit. Compared to the English version of the SAS-A, Spanish version was nearly as good of a better fit. Coefficient alphas were similar to those obtained using the English form of the SAS-A (LaGreca & Lopez, 1998), ranging from .87 to .94 across scales. Authors suggested that this measurement study provides support for the SAS-A to be used with a Spanish-speaking population.

*The Child Behavior Checklist (CBCL).*

The Child Behavior Checklist can be used to assess emotional problems as well as attention and social concerns (Goodman & Scott, 1999). A study by Gross, Fogg, Young, Ridge, Cowell, Richardson, and Silvan (2006) was completed in which the Child Behavior Checklist (CBCL) was completed by parents of two-to-four-year old children who represented a diverse set of races, ethnicities, incomes, and language backgrounds. Overall model fit was assessed through CFA based on the relative chi-square (ratio of the chi-square to the degrees of freedom) and the root-mean-square error of approximation (RMSEA). The authors found that despite language, racial, and socio-economic differences, the model was a good fit when translated to Spanish. The RMSEA statistics were both at .03 and the relative chi-square was 1.66 for the English form and 1.67 for the Spanish form.

## IV. Psychometrics of the sdq across Cultures and Languages

The Strengths and Difficulties Questionnaire (SDQ) was developed in the United Kingdom by Robert Goodman as a rating instrument to assess youth behavior (Goodman, 2001). There are five scales generating scores for Emotional Symptoms (ES), Conduct Problems (CP), hyperactivity-inattention (HI), peer problems (PP), and prosocial behavior (PB), as well as a Total Difficulties (TD) composite score (Goodman, 2001).

Goodman (2001) collected SDQs from parents, teachers, and self-informants in a nationwide epidemiological sample of over 10,000 British students ages 5 to 15. Ninety-six percent of the informants were parents (Goodman, 2001). Internal consistency was assessed and Coefficient alpha coefficients were generally satisfactory for scores representing the five factors, with a mean of .73 across all forms. Table 1 depicts coefficient alpha across subscales for the reviewed SDQ studies. The internal consistency of the TD category was sufficient, with a coefficient alpha of .82. Factor analytic results indicated that all 25 items loaded more heavily onto their respective factors than any of the additional factors. Goodman (2001) noted many items on the HI scale and PP scale on the teacher and self-informant form also substantially loaded (.34 to .52) onto the PB scale. These items were all positively worded indicating a general tendency for positive statements to load onto the PB scale. In addition, the predicted five-first-order factor (5F) structure consisting of the five scales was confirmed.

*Table 1 :* Coefficient Alphas across Strengths and Difficulties Questionnaire Parent-Informant Studies

| | Goodman 2000 | Hawes et al. 2004 | Muris et al. 2002 | Mean |
|---|---|---|---|---|
| Language of Form | English | English | Dutch | |
| Children's Age (years) | 5-15 | 4-9 | 9-15 | |
| Scale | | | | |
| Emotional Symptoms (ES) | .67 | .66 | .70 | .68 |
| Conduct Problems (CP) | .63 | .66 | .55 | .61 |
| Hyperactivity-Inattention (HI) | .77 | .80 | .78 | .78 |
| Peer Problems (PP) | .57 | .59 | .66 | .61 |
| Prosocial Behavior (PB) | .65 | .70 | .68 | .68 |
| Total Difficulties (TD) Composite | .82 | .82 | .80 | .81 |

Hawes and Dadds (2004) analyzed the parent form of the SDQ administered to a large Australian community sample of parents of children ages 4 through 9. Coefficient alpha ranged from .59 to .80. The 5F structure was examined separately for males and females using principal component analyses with oblimin rotation. Results supported the 5F structure, with factor loadings generally stronger for boys than for girls. Consistent with Goodman's study (2001), cross loading occurred with a conduct scale item relating to obedience. Hawes and Dadds (2004) noted that the utility of this item as an indicator of conduct problems may be unreasonable. Using a more negatively worded statement (i.e., "generally disobedient" rather than "generally obedient") may produce a better indicator of conduct problems.

Muris, Meesters, & van den Berg (2002) studied more than 500 parents of children and adolescents using the Dutch form of the SDQ. Internal consistency was generally satisfactory, with a mean coefficient alpha of .70 for scores. However, Coefficient alpha for the CP scale was notably lower (α = .55) compared to the rest of the scales and the TD composite (α ranged from .66 to .80). The five factors (ES, CP, HI, PP, and PB) all had Eigen-values greater than 1.0 (i.e. 4.8, 2.5, 2.0, 1.3, and 1.2). They also accounted for 47.6 percent of the total variance. In addition, all of the items loaded strongly onto their respective factors.

While the aforementioned studies are representative of the large body of research that has been conducted on the SDQ, very little of this research has focused on the preschool version of the measure. In fact, a recent review (Stone, Otten, Engels, Vermulst, & Jannsens, 2010) of 48 research studies on the SDQ included only two studies that extended as young as the three-year-old population, and none focused exclusively on three through five-year-old children, as the current study does. The current study will fill a gap in the research by focusing exclusively on this population.

## V. Research Questions

The current study was inspired by the need for a Spanish language measure of behavior problems from with valid inferences can be drawn, and by the availability of the SDQ in several languages to meet this need. Research questions included:

1. Are there mean differences in SDQ scores based on the language of forms (English versus Spanish)?
2. Are there reliability differences in SDQ scores based on the language of forms (English versus Spanish)?
3. Is the internal structure validity evidence of SDQ scores different based on the language of forms (English versus Spanish)?

## VI. Method

### *a) Participants*

Participants in this study included 363 English-speaking parents and 334 Spanish-speaking parents of preschool age children (ages 3-5) who took part in the California University (Irvine) Initiative for the Development of Attention and Readiness (CUIDAR) program over a four-year period, from 2004-2008. The sample was predominantly Mexican-American (originating from Mexico), regardless of whether the forms were completed in English or Spanish. Both subsamples were well-balanced with regard to gender, and were composed of roughly 1/3 three-year-old children, 1/2 four-year-old children, and 1/6 five-year-old children. The English speaking sample was predominantly Mexican American (43%) and included representative subsamples of European Americans (18%) and African Americans (15%). The Spanish speaking subsample was predominantly Mexican American (85%) and included a representative subsample of Other Hispanic persons (13%). The English speaking parents were more educated on average than the Spanish speaking parents, with about half of the former having completed some college, and about half of the latter not completing high school. Further demographic information is reported in Table 2.

*Table 2 :* Demographic Information across Samples

| | **English Form** | **Spanish Form** |
|---|---|---|
| | **(*n* = 363)** | **(*n* = 334)** |
| Gender | | |
| Female | 45% | 54% |
| Male | 55% | 46% |
| Child's Age | | |
| Three years | 32% | 32% |
| Four years | 53% | 51% |
| Five years | 15% | 18% |
| Child's Ethnicity | | |
| Mexican American | 43% | 85% |
| European American | 18% | 0% |
| African American | 18% | 0% |
| Biracial | 9% | 2% |
| Other Hispanic | 7% | 13% |
| Other NonHispanic | 5% | 0% |
| Parent's Education Level | | |
| Did Not Complete HS | 18% | 50% |
| HS Diploma/GED | 26% | 29% |
| Some College | 49% | 13% |
| Bachelor's Degree | 3% | 6% |
| Advanced Degree | 4% | 2% |

*Note. HS = high school; GED = general equivalency diploma*

CUIDAR is an early intervention program that was designed to reduce potential barriers (e.g., lack of knowledge, lack of insurance, and cultural issues) to screening and intervention for behavioral disorders that may disproportionally affect low-socioeconomic status and minority families (Lakes, Kettler, Schmidt, Haynes, Feeney-Kettler, Kamptner, Swanson, & Tamm, 2009; Lakes, Vargas, Riggs, Schmidt, & Baird, 2011). The goal of CUIDAR is to identify children with attention and behavioral difficulties prior to entering the school system so they will have a more successful educational experience (Lakes et al., 2009). The parent education model used in this program is a modified version of the original Creating Opportunities for Parent Empowerment (COPE) program (Cunningham, Bremner, & Boyle, 1995), which focuses on parent-child interactions, building self-efficacy, and identifying and correcting common parenting errors.

### *b) Measures*

The Spanish, preschool version of the SDQ is used to assess youth ages 3 through 5 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; Goodman, 2001). There are forms for parents, teachers, and self-raters to complete. (Only the parent forms were used in the current study.) The five scales are each based on five items. The TD composite is computed from the four problem scales (i.e., every scale except PB). The theoretical structure of the SDQ is five individual factors representing the five scales. The Spanish version used in the current study is intended to be a direct translation of the English version, with the same factor structure. The Spanish SDQ was used instead of the Spanish (Rio de la Plata) SDQ because the former was more aligned with the Spanish typically spoken in southern California.

### *c) Procedures*

Analyses were conducted using an extant database from the CUIDAR program, and were approved by the institutional review board of the lead author. During the introductory session of CUIDAR, parents were invited to participate in a research study designed to evaluate the effectiveness of the 10-week intervention. As part of their entrance into the research study, participants completed the SDQ. Participants also completed a demographic questionnaire, which included questions regarding race, ethnicity, country of origin, and parent education level. Participants were given an SDQ form in either English or Spanish, based on whether they had self-enrolled in a English- or Spanish-speaking parenting group.

### *d) Data Analysis*

Data were analyzed to determine whether the English and Spanish versions of the SDQ differed with regard to the magnitude of scores and their internal structure. Independent samples t-tests were used to compare mean scores between the two forms at both the subscales and composite level. Reliability was estimated using Coefficient alpha at both the composite and scale levels. CFA was used to examine the internal structure validity evidence.

As part of the CFA, Several indicators were calculated including the normed fit index (NFI), goodness of fit index (GFI), and the comparative fit index (CFI), indicating how well the specific data is structured in relation to the proposed model. The CFI also indicates the fit of a target model to the fit of an independent model, which assumes all variables are uncorrelated (Bentler, 1990). The NFI compares the null model and target model and indicates how well the proposed model improves the fit relative to the independent model (Bentler, 1990). The GFI involves the variances and covariances jointly explained by the model (Joreskog and Sorbom, 1986). All of the aforementioned indices require a statistic of .92 or more to be considered acceptable (Hair Jr. et al., 2010). None of these tests is affected by sample size and normality of distribution.

Other goodness-of fit- statistics used in this study include the Standardized Root Mean Square Residual (SRSMR) and the Root Mean Square Error of Approximation (RMSEA). Following Hair Jr. et al.'s (2010) heuristics for goodness of fit indices, along with our sample size and number of variables, we considered an SRSMR of .08 or less a good fit and an RMSEA of .07 or less a good fit. Akaike's Information Criterion (AIC) was calculated as an indicator of each model's fit relative to its parsimony. Because there are many ways to interpret the findings from CFA, the various multiple fit statistics were considered collectively to represent various perspectives (Campbell, Gillaspy, and Thompson, 1995).

These analyses were used to compare the relative fit of multiple models, including a Five First Order Factor (5F) Model consisting all five scales, a Five First Order within One Second Order Factor (5F1S) model consisting of all five scales scores nested within a second order TD score, and a Four First Order Factors within One Second Order Factor (4F1S) model consisting of the four problem behavior scales nested within the second order TD score and the non-nested PB scale (the 4F1S model is consistent with the SDQ scoring instructions, which indicate TD is the sum of four of the scales).

## VII. Results

Mean ratings of the scales were very similar across the two forms (see Table 3). Mean ratings were significantly higher on the TD scale, t(1.98) = 3.92, $p < .05$, and the HI scale, t(3.47) = 12.04, $p < .01$, when the SDQ was completed in English. Although the difference in mean scores was significant, the effect sizes of the difference between the two forms of the TD scale (d = .24) and HI scale (d = .14) were small. No other differences were significant.

*Table 3 :* Means and Standard Deviations of Parent Ratings across Samples and Scales

| SDQ Scale | **English Form** | **Spanish Form** |
|---|---|---|
| Emotional Symptoms (ES) | 2.19 (2.05) | 2.16 (1.94) |
| Conduct Problems (CP) | 3.63 (2.45) | 3.46 (2.01) |
| Hyperactivity-Inattention (HI) | 4.95* (2.54) | 4.33 (2.20) |
| Peer Problems (PP) | 2.56 (1.90) | 2.51 (1.73) |
| Prosocial Behavior (PB) | 7.25[1] (2.15) | 7.01[a] (2.00) |
| Total Difficulties (TD) Composite | 13.46* (6.43) | 12.52 (5.45) |

*Note: Range of possible ratings is (0-10) on Emotional Symptoms, Conduct Problems, Hyperactivity-Inattention, Peer Problems, and Prosocial Behavior. Range of possible ratings for Total Difficulties is (0-40).* [1] *Higher Ratings are desirable on the Prosocial Behavior Scale.*

** = Significantly higher mean rating on English Form compared to Spanish Form (p < .05).*

Global Journal of Human-Social Science (G) Volume XV Issue I Version I Year 2015

### a) *Reliability*

For the TD scale (English α = .81, Spanish α = .73) and for all five subscales, the coefficient alpha was higher for the score from the English form (see Table 4). On the SDQ English form two of the five scales were in the moderate range, two were in the low range, and one was in the very low range. On the Spanish version of the SDQ, alphas for all five scales were in the very low range.

*Table 4 :* Reliability Coefficients across Forms

| SDQ Scale | English Form | Spanish Form |
|---|---|---|
| Emotional Symptoms (ES) | .65 | .57 |
| Conduct Problems (CP) | .74 | .59 |
| Hyperactivity-Inattention (HI) | .73 | .59 |
| Peer Problems (PP) | .47 | .35 |
| Prosocial Behavior (PB) | .69 | .59 |
| Total Difficulties (TD) Composite | .81 | .73 |

### b) *Confirmatory Factor Analysis*

Six confirmatory factor analyses were performed corresponding to two forms and three models. A comparison of indices across analyses follows.

*English 5 F Model.*

The 5F model for the SDQ in English was a good fit, with the NFI (.88), the CFI (.91), and the GFI (.87) each at or approaching .92. The SRSMR (.07) and RMSEA (.07) also indicated good fit. The 5F model accounted for between 5% and 52% of the variance in each individual item. The saturated model had a lower AIC (650.00) than did the 5F Model (996.12), indicating that the saturated model was a better fit, when not considering theory. The AIC of the Independence model (6385.36) was much higher than either. Table 5 summarizes these indices across forms and models. Factor loadings were high for the CP Factor with four out of five items exceeding .60 and moderately high for the ES, HI, and PB factors. Loadings were lower and more difficult to interpret for the PP Factor. Three of the five items linked to this factor were below .30. Table 6 reports factor loading for each item across forms and models.

*Table 5 :* Goodness of Fit Indices across Models and Forms

| | Englis 5F | h form 5F1S | | | Spanish form | |
|---|---|---|---|---|---|---|
| | | | 4F1S | 5F | 5F1S | 4F1S |
| Indices | | | | | | |
| NFI | .88 | .87 | .84 | .74 | .70 | .68 |
| CFI | .91 | .90 | .88 | .80 | .76 | .74 |
| GFI | .87 | .86 | .85 | .85 | .82 | .83 |
| SRSMR | .07 | .07 | .12 | .08 | .09 | .10 |
| RMSEA | .07 | .08 | .08 | .08 | .09 | .09 |
| AIC | 996.12 | 1077.60 | 1141.45 | 1103.85 | 1270.86 | 1259.50 |

*Note. 5F = Five First Order Factor Model; 5F1S = Five First Order within One Second Order Factor Model; 4F1S = Four First Order within One Second Order Factor Model; NFI = normed fit index; CFI = comparative fit index; GFI = goodness-of-fit index; SRSMR = standardized root mean square residual; RMSEA = root mean square error of approximation; AIC = Akaike's information criterion*

*Table 6 :* Factor Loadings across Models and Forms

| SDQ Scale/Items | English form | | | Spanish fo | | rm |
|---|---|---|---|---|---|---|
| | 5F | 5F1S | 4F1S | 5F | 5F1S | 4F1S |
| *Emotional Symptoms Scale (ES)* | | | | | | |
| Somatic Complaints | .32 | .32 | .32 | .38 | .35 | .36 |
| Worried | .59 | .58 | .59 | .50 | .48 | .48 |
| Unhappy | .60 | .61 | .61 | .55 | .56 | .57 |
| Nervous/Clingy | .46 | .47 | .46 | .45 | .46 | .45 |
| Many fears | .61 | .61 | .61 | .44 | .47 | .46 |
| *Conduct Problems Scale (CP)* | | | | | | |
| Temper tantrums | .60 | .60 | .63 | .39 | .37 | .41 |
| Obedient | .63 | .62 | .57 | .53 | .57 | .46 |
| Fights w/children | .61 | .63 | .61 | .56 | .54 | .58 |
| Lies/Cheats | .64 | .62 | .66 | .48 | .46 | .50 |
| Steals | .52 | .53 | .54 | .35 | .34 | .40 |
| *Hyperactivity-Inattention Scale (HI)* | | | | | | |
| Restless/Overactive | .68 | .67 | .68 | .64 | .61 | .64 |
| Fidgeting/Squirming | .72 | .71 | .72 | .62 | .55 | .62 |
| Distracted | .63 | .63 | .64 | .51 | .51 | .51 |
| Thinks before Acting | .40 | .43 | .40 | .28* | .34 | .27* |
| Attention Span | .52 | .53 | .51 | .28* | .36 | .29* |
| *Peer Problems Scale (PP)* | | | | | | |
| Solitary | .26* | .25* | .28* | .17* | .19* | .22* |
| One good friend | .47 | .45 | .42 | .46 | .42 | .36 |
| Liked by other children | .63 | .63 | .61 | .58 | .57 | .54 |
| Bullied by other children | .23* | .23* | .29* | .19* | .23* | .27* |
| Gets along w/adults more than peers | .29* | .30* | .33 | .13* | .20* | .24* |
| *Prosocial Behavior Scale (PB)* | | | | | | |
| Consider of others | .62 | .65 | .58 | .41 | .41 | .42 |
| Shares | .58 | .58 | .53 | .48 | .43 | .37 |
| Helpful | .51 | .51 | .59 | .47 | .48 | .56 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Kind | .60 | .57 | .53 | .54 | .55 | .48 |
| Volunteers | .49 | .50 | .60 | .49 | .53 | .57 |

*Note: SDQ = Strengths and Difficulties Questionnaire; 5F = Five First Order Factor Model; 5F1S = Five First Order within One Second Order Factor Model; 4F1S = Four First Order within One Second Order Factor Model. * = at or below .30 considered low factor loading.*

*English 5F1S Model.*

The 5F1S model for the SDQ in English was a good fit, with the NFI (.87), the CFI (.90), and the GFI (.86) each approaching .92. The SRSMR (.07) also indicated good fit. The RMSEA (.08) indicated a moderate fit. The 5F1S model accounted for between 5% and 51% of the variance in each individual item. The saturated model had a much lower AIC (650.00) than did the 5F1S model (1077.60), indicating that the saturated model was a better fit, when not considering theory. The AIC of the Independence model (6385.36) was much higher than either. Factor loadings were high for the CP Factor, with four out of five items exceeding .60, and moderately high for the ES, HI, and PB factors. Loadings were lower and more difficult to interpret for the PP Factor. Three of the five items linked to this factor were at or below .30.

*English 4 F 1S Model.*

The 4F1S model for the SDQ in English was a moderate fit, with the NFI (.84), the CFI (.88), and the GFI (.85) each exceeding .80. The SRSMR of (.12) and RMSEA (.08) indicated moderate fit. The 4F1S model accounted for between 8% and 53% of the variance in each individual item. The saturated model had a much lower AIC (650.00) than did the 4F1S model (1141.45), indicating that the saturated model was a better fit, when not considering theory. The AIC of the Independence model (6385.36) was much higher than either. Factor loadings were high for the CP Factor, with three out of five items exceeding .60, and moderately high for the ES, HI, and PB factors. Loadings were again lower and more difficult to interpret for the PP factor.

*Spanish 5 F Model.*

The 5F model for the SDQ in Spanish was a moderate fit, with the NFI (.74), the CFI (.80), and the GFI (.85) each at or approaching .80. The SRSMR (.08) indicated good fit. The RMSEA (.08) indicated a moderate fit. The 5F Model accounted for between 2% and 38% of the variance in each individual item. The saturated model had a significantly lower AIC (650.00) than did the 5F Model (1103.85), indicating that the saturated model was a better fit, when not considering theory. The AIC of the Independence model (3330.00) was much higher than either. Factor loadings were moderate for the CP, ES, and PS factors. Loadings were lower and more difficult to interpret for the PP and HI factors. Three of the five items linked to the PP Factor were below .30. Although two items associated with the HI Factor loaded highly onto their factor, two of the loadings were below .30.

*Spanish 5 F 1S Model.*

The 5F1S model for the SDQ in Spanish was a poor fit, with the NFI (.70), the CFI (.76), and the GFI (.82) far below .92. The SRSMR (.09) and RMSEA (.09) both indicated moderate fit. The 5F1S model accounted for between 2% and 46% of the variance in each individual item. The saturated model had a much lower AIC (650.00) than did the 5F1S model (1270.86), indicating that the saturated model was a better fit, when not considering theory. The AIC of the Independence model (3330.00) was much higher than either. Factor loadings were moderate for the CP, ES, and PS factors. Loadings were lower and more difficult to interpret for the PP and HI factors. Three of the five items linked to the PP Factor were below .30. Although two items associated with the HI Factor loaded highly, two of the loadings were only slightly above .30.

*Spanish 4 F 1S Model.*

The 4F1S model for the SDQ in Spanish was a poor fit, with the NFI (.68), the CFI (.74), and the GFI (.83) far below .92. The SRSMR (.10) and RMSEA (.09) both indicated moderate fit. The 4F1S model accounted for between 5% and 41% of the variance in each individual item. The saturated model had a much lower AIC (650.00) than did the 4F1S model (1259.50), indicating that the saturated model was a better fit, when not considering theory. The AIC of the Independence model (3330.00) was much higher than either. Factor loadings were moderate for the CP, ES, and PS factors. Loadings were lower and more difficult to interpret for the PP and HI factors. Two of the five items linked to the PP Factor were below .30. Although two items associated with the HI Factor loaded highly, two of the loadings were below .30.

*English versus Spanish Models.*

Data from the English forms fit the models better than did data from the Spanish forms. The average NFI, CFI, and GFI for the English models were all substantially higher than averages for the Spanish models. The averages of the Standardized RMRs (.09) and RMSEAs (.08) were identical across models. Models in both English and Spanish accounted for approximately the same percentage (2% to 50%) of the variance in each individual item. Factor loadings were much higher across the English models than across Spanish models. Items on the English models loaded highly onto the CP Factor, and moderately onto the EP, HP, and PB factors. Items did not load well onto the PP Factor. Factor loadings were moderate, at best, for the Spanish models. Similar to the English models, items related to being solitary, getting bullied, and relating better with adults than children loaded poorly onto the PP Factor. Unique to the Spanish model, loadings were inconsistent on the HI Factor.

## VIII. Discussion

This study contributes important information regarding the reliability and validity of scores derived from the SDQ Spanish version for parents of preschoolers. Parent raters who took part in CUIDAR assessed their preschool age children's behaviors using the SDQ as part of their entrance into the intervention program. In this study, the psychometric properties of scores were assessed in order to explore mean rating differences between the English and Spanish versions of the SDQ, along with coefficient alpha indicators of reliability at the scale and composite level, and internal structure validity evidence. Results indicated scale mean scores were very similar across both forms of the SDQ. Reliability coefficients indicated alphas were higher for scores obtained on the English form compared to the Spanish form. Finally, the 5F Model that is predominant in the literature was the best-fit and most valid representation of all 25 items of the SDQ, regardless of the language of the form. The 5F1S model was comparable in English, and the 4F1S model that is consistent with SDQ scoring instructions was the worst fit regardless of form. The English form yielded data that fit better across models than did the Spanish form.

### *a) Group Differences*

The first research question was whether there are mean differences in SDQ scores for students from Spanish-speaking families versus students from English-speaking families. Mean ratings were similar across English and Spanish forms, with significant but small mean differences on the HI scale and the TD composite. The finding that these differences were small is supportive of the SDQ, indicating that it is not systematically biased to produce higher scores when used with either population.

### *b) Precision of Measurement*

The second research question addressed how well the items from the two forms fit together to yield scale scores. Coefficient alphas for scores on the SDQ scales were compared at the scale and composite levels. Alphas were higher across scales on the English form of the SDQ, compared to the Spanish form. The TD scores in English were high enough to make low stakes decisions, or to be included as one of multiple measures in a thorough assessment. The score reliabilities were not high enough for making critical clinical or educational decisions.

Prior research has yielded similar reliability coefficients at the scale and composite level. Goodman (2001) found coefficient alphas in the low to moderate range, with only the TD composite in the good range. Hawes et al. (2004) and Muris et al., (2002) obtained similar results, with alphas ranging from the low to moderate range at the scale level, and above .80 and in the good range for the TD scale. It is difficult to obtain alphas in the adequate or good range when there are only five items on each scale. Although a benefit of the SDQ is its brevity, increasing the number of items could make scores more reliable.

### *c) Internal Structure Validity Evidence*

The third research question involved whether the factor structure of the SDQ in Spanish differed from the factor structure of the SDQ in English. Three factor models were evaluated through CFA on both the English and Spanish forms of the SDQ. The first was a 5F Model, which has been confirmed in prior literature to fit. It consists of five factors from which scale scores are yielded: ES, CP, HI, PP, and PB. The second model evaluated was a 5F1S model with all factors nested within the TD factor. The third model evaluated was a 4F1S model with four factors nested within the TD composite, isolating the PB factor, as is implied by the SDQ scoring instructions.

In this study, regardless of whether the form was completed in English or Spanish, the 5F Model was the best fit and most valid representation of the 25 items on the measure. Factor loadings were consistently higher from the English forms compared to the Spanish forms. However, across models and forms, loadings were consistently very low for items on the PP Scale. This may be due to some items within this index being reverse scored and others being scored normally. Having a more uniform scoring system within the index would likely yield higher loadings.

Prior research has consistently indicated that the 5F Model is a good fit. Similar to this study, Goodman (2001) confirmed the 5F Model and indicated that all 25 items loaded onto their intended factors. Hawes and Dadds (2004) also confirmed the 5F Model with parents of Australian children, ages four through nine. They found that factor loadings were generally stronger for boys than for girls, but that the design was a good fit regardless of gender.

A strength of the current study is that CFA was used with multiple models. Prior studies, which assessed the factor structure of the SDQ, did not do this. Goodman (2001), Hawes and Dadds (2004), and Muris, Meesters, & van den Berg (2002) all confirmed the 5F model of the SDQ, but did not include comparison with other models. For the English form, the 5F1S was a comparable model to the 5F, providing some evidence for pooling the scale scores into a TD composite. This model faired better than did the 4F1S that is implied by the scoring instructions, which do not include the PB in calculation of the TD. These findings indicate that, when using the English form, a method that calculates a TD score from all five subscales might be superior. For the Spanish form, neither the 5F1S model nor the 4F1S model fit the data well.

Regardless of model, the internal structure evidence for the Spanish form was inadequate and inferior to the evidence for the English form. Similar to findings obtained when using the Spanish form of the BERS-2, these results indicate that the properties of the SDQ are negatively altered through the translation process (Sharkey et al., 2009). Coupled with the findings on reliability, these results indicate that the Spanish form of the SDQ might be revised and further evaluated before being used in educational or clinical settings to measure or identify behavioral problems in preschool children. The findings also reinforce that whenever possible, researchers should evaluate and report on the reliability and validity of scores obtained in their research, rather than relying solely on prior measurement studies (e.g., Yin & Fan, 2000; Lakes, 2012).

### *d) Implications for Practice*

When using the SDQ for a preschool, Spanish-speaking, Mexican-American population, the current findings indicate that a conservative decision rule should be used. This recommendation is based on the TD score being lower on average, and the reliability and validity evidence being poorer, compared to the evidence for the English form. Collectively, these results indicate that scores from the Spanish form will be lower, and that error will be contributing to more of their variance. Therefore, difficulties will be harder to detect (i.e., less likely to be manifested in high scores). If the Spanish form of the SDQ is used for a low stakes purpose (e.g., identification for a group behavioral program), a lower cut score might be considered. However, it is always preferable to use a measure that yields more reliable scores from which more valid inferences can be made, and the current study provides no support for using the Spanish form of the SDQ for high stakes decisions.

Depending on the specific type of behavior problem for which one is screening, other measures such as the SAS-A and CBCL have been shown to produce scores with acceptable psychometric properties in their Spanish versions. Compared to the SDQ, the SAS-A when translated still produces scores that demonstrate good reliability and internal structural validity. However, it is not as similar to the SDQ as one would like because it can only be used in an adolescent population with self-raters.

The CBCL is another measure that can be used for many of the same purposes as the SDQ (Goodman & Scott, 1999). The CBCL is widely used in schools and has good psychometric properties in its Spanish translated version. Furthermore, as mentioned earlier, studies have found that the SDQ and CBCL are comparable in many ways. The two measures correlate highly, address similar behaviors, and discriminate between low and high-risk populations (Goodman & Scott, 1999). Therefore, the CBCL in Spanish may be preferred to the SDQ in Spanish, for preschool Mexican-American children.

### *e) Limitations*

The generalizability of these findings is limited in several ways. The SDQ has forms for children up to age 16; however this study is limited in that only children 3 through 5 were rated. Mean ratings may have differed if the sample represented a larger age range of students, and prior research has demonstrated that restriction of range in a study sample can reduce the observed reliability of scores (Henson, Kogan, & Vacha-Haase, 2001; Lakes, 2012). Also, there was an unequal distribution of ethnicities represented in this sample, with Mexican-American children being the most highly represented. It is unknown how generalizable the results of this study would be in communities where the Mexican-American population is not as high. The most conservative interpretation would be that the results are only generalizable to the Spanish-speaking population of southern California. While it is likely that results would be similar for many surrounding areas in California, less is known about the generalizability of the findings to Spanish speaking populations from cultures and geographical regions not represented in this sample. Lastly, the current study did not include any measure of acculturation, which could be a confounding variable when looking at the psychometrics of an instrument across forms defined by language.

### *f) Future Research*

Future studies regarding the SDQ could analyze changes in mean ratings as children grow older. In this study, SDQ ratings were only taken at the point of entry into the CUIDAR program. It would be helpful to examine how ratings may change over time as children develop.

Similarly, it would be interesting to interpret what similar ratings over time may indicate about the stability of problems or areas of strength that youth possess.

Another area of research could involve examining mean parent ratings of the English and

Spanish forms of the SDQ, using groups of parents born in Mexico and born in the United States, in order to analyze whether country of origin impacts the relationship between language and psychometrics of the SDQ. This design could also be expanded to other counties.

Research could also be focused on improving the SDQ at the item level. One might consider comparing the standard version of the measure which has three-point item level response choices with versions that have four or five levels of response. It is possible that the latter would have better psychometric properties.

Finally, the factor structure of the SDQ should be evaluated in all of the languages into which the measure has been translated. Doing so would indicate whether the translation of the SDQ items into different languages has resulted in changes in psychometric properties.

## IX. Conclusions

As part of their entrance into CUIDAR, parent raters assessed their preschool age children's behaviors using the SDQ. Data was collected over a four-year period, from 2004-2008. In this study, the psychometric properties of scores were assessed in order to explore mean rating differences between the English and Spanish versions of the SDQ, along with coefficient alpha indicators of reliability at the scale and composite level, and factor structure differences. Results indicated that mean ratings of the individual scales and the TD scales were very similar across both forms of the SDQ. Reliability coefficients indicated alphas were higher for the English form compared to the Spanish form at the scale and composite levels. On the TD composite, there was good reliability when the form was completed in English. Finally, the 5F Model was the best-fit and most valid representation of the 25 items of the SDQ, despite the language of the form. The 5F1S model was also a good fit for the English form, but not for the Spanish form. The English form yielded data that fit better, compared to that yielded by the Spanish form, regardless of model. Thus, it is important for practitioners to utilize caution when using the SDQ in a Spanish-speaking, Mexican-American population of preschool children.

## References Références Referencias

1. Achenbach, T.M. (1991). *Manual for the Child Behavior Checklist and 1991 Profile*. Burlington, VT: University of VT, Department of Psychiatry.
2. American Educational Research Association. (1999). *Standards for Educational and Psychological Testing.* American Educational Research Association, American Psychological Association, National Council on Measurement in Education.
3. Washington, DC: American Educational Research Association. Bentler, P.M (1990). Comparative fit indices in structural models. *Psychological Bulletin,*107, 238-246.
4. Buckley, J.A., Ryser, G., Reid, R., & Epstein, M.H. (2006). Confirmatory factor analysis of the Behavioral and Emotional Rating Scale-2 (BERS-2) Parent and Youth Rating Scales. *Journal of Child and Family Studies,* 15, 27-37.
5. Campbell, T.C., Gillaspy, J.A., & Thompson, B. (1995, January). The factor structure of the Bem Sex-Role Inventory (BSRI): A confirmatory factor analysis. Paper presented at the annual meeting of the Southwest Educational research Association, Dallas. (ERIC Document Reproduction Service No. ED 380 491)
6. Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology,* 78, 98- 104.
7. Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. J*ournal of Child Psychology and Psychiatry,* 38, 581-586.
8. 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.
9. Goodman, R., Ford, T., Simmons H., Gatward, R., & Meltzer, H. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry,* 177, 534-539.
10. Goodman R, Scott S (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: Is small beautiful? *Journal of Abnormal Child Psychology,* 27, 17-24.
11. Gross, D., Young, M., Fogg, L., Ridge, A., Cowell, J., Richardson, R.,& Sivan, A. (2006).
12. The equivalence of the child behavior checklist/11/2-5 across parent race/ethnicity, income level, and language. *Psychological Assessment, 18*, 313-323.
13. Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6,* 427-439.
14. Hair, J., Black, W., Babin, B., Anderson, R. (2010). *Multivariate Data Analysis.* Upper Saddle River, NJ: Prentice Hall.
15. Hawes DJ, Dadds MR (2004). Australian data and psychometric properties of the Strengths and Difficulties Questionnaire. *Australian and New Zealand Journal of Psychiatry,* 38, 644-651.
16. Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement,* 61, 404-420.

17. Hopkins, W. G. (2002). A scale of magnitudes for effect statistics. *A New View of Statistics.* Retrieved October9,2011,fromhttp://sportsci.org/resource/stats/effectmag.html
18. Felt, L. & Seonghoon, K. (2006). Testing the difference between two alpha coefficients with small samples of subjects and raters. *Educational and Psychological Measurement,* 66, 589-600.
19. Jimerson, S. R., Sharkey, J. D., Nyborg, V. M., & Furlong, M. J. (2004). Strength-based assessment and school psychology: A summary and synthesis. *California School Psychologist, 9*, 9-20.
20. Joreskog, K.G. & Sorbom, D. (1986). Lisrel VI. 4th Ed. Mooresville, Indiana: Scientific Software, Inc.
21. Kettler, R.J., Elliott, S.N., Beddow, P.A., Compton, E., McGrath, D., Kaase, K., Bruen, C.,Ford, L., & Hinton, K. (2010). What does an alternate assessment measure? A multitrait-multimethod analysis. *Exceptional Children, 76* (4), 457-474.
22. La Greca, A.M. & Lopez, N. (1998). Social anxiety among adolescents: Linkage with peer relations and friendships. *Journal of Abnormal Child Psychology,* 26, 83-94.
23. Lakes, K.D. (2012, in press). Restricted sample variance reduces generalizability. *Psychological Assessment.*
24. Lakes, K. D., Lopez, S., & Garro, L. (2006). Cultural competence and psychotherapy: Applying anthropologically informed conceptions of culture. *Psychotherapy Theory, Research, Practice, and Training*, 4, 380-396.
25. Lakes, K.D., Kettler, R.J., Schmidt, J., Haynes, M., Feeney-Kettler, K.A., Kamptner, L., et al. (2009). The CUIDAR early intervention parent-training program forpreschoolers at risk for behavioral disorders: An innovative practice for reducing disparities in access to service. *Journal of Early Intervention,* 31, 167-178.
26. Running head: INTERNAL STRUCTURE OF THE SDQ-SPANISH 27Lakes, K.D., Vargas, D.,* Riggs, M., Schmidt, J.,* & Baird, M.* (2011). Parenting intervention to reduce attention and behavior difficulties in preschoolers: A CUIDAR evaluation study. *Journal of Child and Family Studies, 20, 648-659.*
27. Muris, P., Meesters, C., & Van den Berg, F. (2003). The Strengths and Difficulties Questionnaire (SDQ): Further evidence for its reliability and validity in a community sample of Dutch children and adolescents. *European Child and Adolescent Psychiatry,* 12, 1-8.
28. Olivares, J., Ruiz, J., Hidalgo, M., Garcia-Lopez, L., Rosa, A., & Piqueras, J.(2005). Social anxiety scale for adolescents (SAS-A): psychometric properties in a Spanish-speaking population. *International Journal of Clinical and Health Psychology, 5,* 85-97.
29. Pedrotti, J. T., & Edwards, L. M. (2010). The intersection of positive psychology and multiculturalism in counseling. In J. G. Ponterotto, J.M. Casas, L. A. Suzuki, & C. M. Alexander (Eds.), Handbook of multicultural counseling (3rd ed.,pp. 165–174). Thousand Oaks, CA: Sage
30. Sharkey, J. D., You, S., Morrison, G. M., & Griffiths, A. J. (2009). Behavioral and Emotional Rating Scale-2 Parent Report: Exploring a Spanish Version with At-Risk Students. *Behavioral Disorders, 35* (1), 53-65.
31. Smedje, H., Broman, J.E., Hetta, J., Von Knorring, A.L. (1999). Psychometric properties of a Swedish version of the "Strengths and Difficulties Questionnaire". *European Child and Adolescent Psychiatry,* 8,63-70.
32. Smokowski, P.R., Reynolds, A.J., & Bezruczko, N. (1999). Resiliency and protective Running head: INTERNAL STRUCTURE OF THE SDQ-SPANISH 28 factors in adolescence: An autobiographical perspective from disadvantaged youth. *Journal of School Psychology, 37* (4), 425-448.
33. 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, 197-208.
34. Tinkler, B. (2002). *A review of literature on Hispanic/Latino/a parent involvement in K- 12 education.* Retrieved October 28, 2011, from http://www.eric.ed.gov/PDFS/ED469134.pdf
35. Yin, P. & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory Scales: Reliability generalization across studies. *Educational and Psychological Measurement, 60,* 201-223.
36. Youth in Mind Organization. (2012). *SDQ Info.* RetrievedNovember20,2011,fromhttp://www.sdqinfo.com/Running head: INTERNAL STRUCTURE OF THE SDQ-SPANISH 29