Measuring Functional Impairment in Children and Adolescents: Psychometric Properties of the Columbia Impairment Scale (CIS)

Brandon K. Attell[1], Christopher Cappelli[2], Brigitte Manteuffel[1], and Hongli Li[3]

Abstract

The role of measuring functional impairment holds an important place in research, clinical practice, and service provision for children and adolescents. Responding to the growing need to measure serious emotional disturbances at the local, state, and national level, the Columbia Impairment Scale (CIS) was developed in the early 1990s and has remained one of the several popular scales for assessing functional impairment. However, despite the growing popularity of the instrument in research and practice, only a few studies to date have specifically examined the psychometric properties of the CIS. In this article, we describe the results of the first item response theory analysis of the CIS utilizing nationally representative data from the Medical Expenditure Panel Survey ($N = 69{,}966$). The results of our analysis lend support to the essential unidimensionality of the CIS and demonstrate that the scale is most reliable for those who exhibit high levels of functional impairment. Given the psychometric properties of the scale identified by our analysis, we contend that the CIS is a viable measure in the ongoing efforts to establish a national epidemiologic surveillance system to track the prevalence and impact of serious emotional disturbances in children and adolescents.

Keywords

Columbia Impairment Scale, functional impairment, psychometrics, item response theory, behavioral health

Over the past 40 years, the measurement of child and adolescent functional impairment has been a growing topic of interest among service providers, policy makers, researchers, and funders of behavioral health services. In the context of diagnosis, functional impairment is defined as the extent to which presenting symptoms impact an individual's adaptive capacity to function across multiple contexts such as home, school, work, or with other individuals including parents, siblings, or friends (Bird & Gould, 1995; Canino, Costello, & Angold, 1999). With the implementation of the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III*) in 1980, the presence of functional impairment was added as a necessary criterion to render a diagnosis for many mental disorders (Canino et al., 1999). Despite the addition of the impairment criterion in *DSM-III* and its continuation through *DSM-IV* and *V*, what constitutes impairment is largely left to the interpretation of clinicians, although the *DSM* provided general recommendations for assessing impairment using scales such as the Global Assessment of Functioning for *DSM-IV* and the World Health Organization Disability Assessment Schedule for *DSM-V* (Canino et al., 1999; Gold, 2014). Nevertheless, the addition of the impairment criterion as a threshold for diagnosis raised awareness in the service provision and research communities regarding the need for

psychometrically sound instruments that measure functional impairment (Canino, Fisher, Alegria, & Bird, 2013; Gordon et al., 2006; Klein, Dougherty, & Olino, 2005; McMahon & Frick, 2005).

In the health policy arena, the need for instruments that appropriately measure functional impairment was formalized with the passage of the Alcohol, Drug Abuse and Mental Health Services Administration (ADAMHA) Reorganization Act of 1992. ADAMHA created the Community Mental Health Service Block Grant and the Comprehensive Community Mental Health Services Program for Children and Their Families, effectively establishing block grant and demonstration program-funding mechanisms aimed at distributing federal resources to states based on the prevalence of children and adolescents living with serious emotional disturbances (SED;

[1]Georgia Health Policy Center, Georgia State University, Atlanta, GA, USA
[2]Center for Education Integrating Science, Mathematics, and Computing, Georgia Institute of Technology, Atlanta, GA, USA
[3]Department of Educational Policy Studies, Georgia State University, Atlanta, GA, USA

Corresponding Author:
Brandon K. Attell, Georgia Health Policy Center, Georgia State University, P.O. Box 3992, Atlanta, GA 30302, USA.
Email: battell1@gsu.edu

Canino, 2016; Holden et al., 2003). With the creation of these funding mechanisms, ADAMHA federally mandated that states engage in epidemiologic surveillance to track the prevalence of children and adolescents living with SED, requiring that functional impairment be measured to distinguish between symptomology and clinical diagnostic threshold based on the functional impairment criterion of the *DSM* (Ringeisen et al., 2017). As funding was targeted toward programs to address SED in children and adolescents, policy makers expected widespread improvement in functional impairment as a system-level outcome (Brannan, Brashears, Gyamfi, & Manteuffel, 2012). As states and communities responded to these mandates and systematically began to track SED, it was discovered that functional impairment is one of the strongest predictors of mental health service utilization (see Burns et al., 1995; Merikangas, Bromet, & Druss, 2017; National Academies of Sciences, Engineering, and Medicine, 2016).

Against the backdrop of these issues, the Columbia Impairment Scale (CIS; Bird, Shaffer, Fisher, & Gould, 1993) was developed in the early 1990s among a series of other instruments to measure functional impairment (Winters, Collett, & Myers, 2005). The CIS was developed in response to the need for a brief scale capable of lay interviewer or self-report administration. Prior to the CIS, most measures designed to assess functional impairment required the scoring to be conducted by an individual with clinical expertise (Bird et al., 1993). The initial pilot of the CIS was conducted on a sample of 182 children and parents in New York who were enrolled in a community-based epidemiological study examining the utilization of various instruments to assess the prevalence of mental disorders.

Thirteen items, each having categorical response options with zero as no problem at all to four as a very bad problem, were designed to measure functional impairment in four domains: interpersonal relations, broad psychopathological domains, functioning in school or at work, and use of leisure time. Additionally, both a parent and child/youth version of the CIS were piloted for feasibility. The initial psychometric testing (Bird et al., 1993, 1996) included factor analyses to determine the dimensionality of the CIS as well as estimates of internal reliability and test–retest reliability over a two-point time period averaging 15 days. The validity of the CIS was examined by correlating total CIS scores with the Children's Global Assessment Scale (CGAS), also designed to measure functional impairment. Finally, a receiver operator characteristic (ROC) curve was employed comparing the CIS with the CGAS to examine sensitivity and specificity of the instrument and to determine an optimal cut point that demonstrated significant functional impairment for those in the pilot sample.

The initial psychometric analyses of the CIS established the parent version as reputable for utilization in future research and clinical practice. Although the CIS items were originally designed to tap into functional impairment in the four domains described above, the factor analyses carried out by Bird, Shaffer, Fisher, and Gould (1993) indicated that the measures most strongly loaded onto one dominant factor, which they termed

global functional impairment. Internal reliability of the parent and child/youth CIS was respectable across multiple time points, although the α for the child/youth measure was somewhat lower (ranging from .70 to .78) compared to the parent version (ranging from .85 to .89). Similarly, the CIS was well correlated with other indicators of psychological dysfunction as measured by the CGAS, although correlations were lower for the child/youth CIS. The ROC analysis indicated that if total scores for all 13 items were summed, values above 15 represented a clinical indication of functional impairment (Bird et al., 1993, 1996). The consistently lower reliability and validity of the child/youth CIS, in addition to children's limitations in assessing their own cognitive processes (Klein et al., 2005; Ringeisen et al., 2017), resulted in more widespread adoption of the parent CIS in research and clinical practice (Singer, Eack, & Greeno, 2011).

Functional impairment in general was a critical outcome tracked alongside the rising implementation of system of care philosophy (see Manteuffel, Stephens, & Santiago, 2002; Manteuffel, Stephens, Sondheimer, & Fisher, 2008; Walrath et al., 2003), and the CIS in particular has been frequently employed to track changes in impairment at the local, state, and national level among children and adolescents receiving system of care services (Brennan, Nygren, Stephens, & Croskey, 2016; Snyder et al., 2012; Starin et al., 2014; Vishnevsky, Strompolis, Reeve, Kilmer, & Cook, 2012). Compared to other instruments such as the CGAS and Child and Adolescent Functional Assessment Scale (CAFAS), the CIS has frequently been employed because of self-administration capabilities and short completion time (Winters et al., 2005). Additionally, the CIS has been identified as one of the several common scales implemented in most research related to children and adolescents funded by the National Institute of Mental Health (Barch et al., 2016).

The CIS has also exhibited consistent use in clinical practice. For example, many communities and states providing wraparound care planning and management to children and adolescents regularly utilize the CIS to determine eligibility for services (e.g., see Painter, 2012), and many funders of these services require documented evidence of functional impairment for service reimbursement (Canino, 2016; Canino et al., 2013). The CIS has also been instrumental in examining sex and race disparities in access and utilization of mental health services (Garland et al., 2005; Wu et al., 1999; Yeh, McCabe, Hough, Dupuis, & Hazen, 2003) and has recently been utilized to examine trends in behavioral health care among adolescents at the population level (Olfson, Druss, & Marcus, 2015). Additionally, researchers have incorporated the CIS as a control measure for functional impairment in examining factors influencing various behavioral health issues including help-seeking behavior (Gould, Munfakh, Lubell, Kleinman, & Parker, 2002), antisocial behaviors (Bird et al., 2001), childhood executive functioning (Miller & Hinshaw, 2010), and bullying (Klomek et al., 2011).

The use of the CIS across these settings is predicated on the assumption that the scale has strong psychometric properties.

However, even with the uptake of the CIS in research, clinical practice, and systems transformation, the measurement characteristics of the scale have been understudied relative to other instruments measuring functional impairment (Winters et al., 2005). Beyond its initial pilot testing and development, only four studies to date have specifically examined the measurement characteristics of the CIS. Singer, Eack, and Greeno (2011) conducted exploratory and confirmatory factor analyses on the CIS in a sample of 280 mothers whose children were receiving community mental health services in three Pennsylvania communities, finding evidence that in their sample, a 3-factor model best fit the data. On this basis, they called into question the unidimensionality of the CIS, arguing that the CIS measures three latent constructs of impairment at school/work, in socializing, and at home/family (Singer et al., 2011).

Support for the criterion-related validity of the CIS was provided by Steinhausen and Winkler Metzke (2001), who found total CIS scores among a sample of 1,089 youth in one region of Switzerland to significantly correlate with the CGAS, which also measures functional impairment. However, in contrast to these findings, Zielinski, Wood, Renno, Whitham, and Sterling (2014) found little support for criterion-related validity but instead found the CIS to correlate well with scales measuring unrelated constructs such as anxiety and various externalizing and internalizing behaviors in their sample of 77 adolescents with autism spectrum disorder. Finally, using a convenience sample of 180 adolescent students in Italy, Zanon, Tomassoni, Gargano, and Granai (2016) examined test–retest reliability of the four CIS subscales, finding intraclass correlation coefficients ranging from .42 to .52, indicating moderate test–retest reliability (Koo & Li, 2016).

It is worth noting that beyond these four studies specifically examining measurement properties of the CIS, a handful of other analyses have by proxy examined the scale's criterion-related validity by correlating scores on the CIS with instruments being pilot tested or further validated. For example, scores on CIS have been correlated with the CAFAS (Ezpeleta, Granero, de la Osa, Domenech, & Benillo, 2006), the Pediatric Symptom Checklist (Gardner, Lucas, Kolko, & Campo, 2007), the Personal Adjustment and Role Skills Scale (Harris, Canning, & Kelleher, 1996), the Behavioral and Emotional Rating Scale for Youth (Lambert et al., 2015), and the Child Behavior Checklist (Gardner et al., 2007; Harris et al., 1996). In all cases, the CIS demonstrated moderate to strong correlations with other instruments designed to measure impairment or related constructs, further lending support to the validity of the CIS. Kramer et al. (2004) examined parent and youth interrater reliability of several questions taken from the CIS infused into a larger study, finding that the items related to relationships with father, mother, and siblings, as well as involvement in sports or hobbies, demonstrated weak but statistically significant interrater reliability. Finally, community- and population-based studies utilizing the CIS as a control variable have generally reported high internal reliability of the CIS, with Cronbach's α generally reported above .80 (e.g., see Olfson

et al., 2015; Starin et al., 2014; Yeh, Hough, McCabe, Lau, & Garland, 2004).

Taken together, these studies highlight several patterns related to previous research examining the psychometric properties of the CIS. First, there is conflicting evidence that the scale is unidimensional, or that it demonstrates the essential unidimensionality characteristic assumed with many psychological instruments (Slocum-Gori, Zumbo, Michalos, & Diener, 2009). Second, the psychometric properties of the CIS have primarily been examined in several community-based convenience samples with the parents of youth who were enrolled in or seeking mental health services or involved in other public service sectors. As others have pointed out (Hammond, 2006; Streiner, Norman, & Cairney, 2015; Winters et al., 2005) once pilot studies have examined measurement characteristics of an instrument, it is useful to assess the scale considering data from more general, representative samples when possible. Finally, apart from Kramer et al. (2004), the existing research to date on the measurement properties of the CIS has mostly focused on the reliability or validity of the instrument as a whole and has ignored the role that each item plays in shaping the reliability of the instrument. If the CIS is indeed a unidimensional measure, it is important to examine the reliability of each item, as well as the total scale, across the continuum of the underlying impairment trait being measured. In this study, we address these gaps by performing item-level analyses utilizing item response theory (IRT) modeling on data from a nationally representative sample of noninstitutionalized Americans. To the best of our knowledge, this study is the first and only IRT analysis of the CIS.

Method

The data for this study come from the Medical Expenditure Panel Survey (MEPS; Agency for Healthcare Research and Quality, 2017). The MEPS began in 1996 as a set of large-scale surveys of individuals and families, their medical providers, and their employers and currently has data available through 2015. The MEPS routinely collects data regarding health services utilization in the United States including their frequency of use, service costs, and payment methods. The MEPS has two major components, the household component and the insurance component. The measures used in this study are drawn from the household component of the MEPS, which implemented the CIS in 1996 and administered the questions to respondents annually through 2015. These data are considered nationally representative, drawing from a subsample of the National Health Interview Survey.

To capitalize on the availability of the MEPS' CIS data collected over such a large period of time, we pooled the 1996—2015 MEPS household component files and restricted the analysis to those individuals who were presented the questions and had nonmissing responses to all 13 items on the CIS. Due to the longitudinal nature of the MEPS survey design, participants were administered the CIS on two separate occasions. To avoid the correlation of residual measurement error

associated with longitudinal survey data (Hoffman, 2015), this analysis was further restricted to only the first set of CIS responses for each participant. The final sample size for the analysis was 69,966 individuals.

Data Analysis

The psychometric analysis of the CIS broadly followed the steps outlined by Toland (2014). First, drawing on classical test theory (CTT), descriptive statistics were examined for all 13 items including an examination of the percentage of responses that fell within each of the five categories for every item. Internal reliability was also examined (Cronbach's α) as well as item–test, item–rest, and correlations among all CIS items. Next, an item-level analysis was performed on the CIS using IRT modeling. Two key assumptions of IRT are unidimensionality of the measured trait and local independence (Nguyen, Han, Kim, & Chan, 2014). To assess these assumptions, an exploratory factor analysis (EFA) was performed. First, principal axis factoring (PAF) was used to test the number of factors that underlie the data. Then, a varimax rotation was used to obtain a unique factor solution.

The following criteria guided the factor analysis: The Kaiser (1960) rule, which recommends retaining only those factors having Eigenvalues greater than 1, and an examination of the scree plot were used to determine the number of factors present in the CIS data (Cattell, 1960). Next, the factor loadings of items onto the established latent factors were examined, and items with communality values less than or equal to .20 were considered for removal from the analysis (Child, 2006). The root mean square error of approximation (RMSEA) was used as the primary measure for model fit, where values of .06 or lower indicate adequate model fit (Hu & Bentler, 1999). In addition, the R^2 value, which is comparable to the goodness of fit index and the adjusted goodness of fit index, indicates acceptable model fit for values greater than .90. Given essential unidimensionality, the assumption of local independence can also be considered to be satisfied. When the assumption of local independence is met, there are no dependencies among test items other than those attributable to respondents' latent ability (Kolen & Brennan, 2014). Therefore, if the assumption of unidimensionality is satisfied, there are no other latent factors that may influence an examinee's response on other test items, and local independence can also be assumed.

Once it was determined that the data were unidimensional and locally independent, the IRT model was fit to the data. Due to the ordinal nature of the CIS items, a graded response model (GRM; Samejima, 1969) was estimated. The GRM utilizes a maximum likelihood estimation procedure to predict responses to ordinal scale items and is parameterized as:

$$\Pr(Y_{ij} \geq k | a_i, b_i, \theta_j) = \frac{\exp\{a_i(\theta_j - b_{ik})\}}{1 + \exp\{a_i(\theta_j - b_{ik})\}}, \theta_j \sim N(0, 1),$$

where the probability of responding in a particular item category k is conditioned on an item's discrimination (a_i), item's difficulty (b_i), and the latent trait of the individual (θ_j), assumed to be normally distributed with a mean of 0 and standard

deviation of 1. In the case of the GRM, higher discrimination values represent a stronger relationship between the item and the underlying construct measured by the scale. Each difficulty parameter is placed on the z-score scale and represents the point along the latent trait continuum needed for an individual to have a 50/50 chance of responding in that category (Hays, Morales, & Reise, 2000).

Based on the results of the GRM, we performed a follow-up supplementary analysis to examine item characteristics and overall model fit of the CIS if some of the categories were collapsed. Our decision on which categories to collapse was based on the research of Goudie, Havercamp, Jamieson, and Sahr (2013). In their analysis of differences in functional impairment among children whose siblings were either typically developing or had a disability, they found low participant responses for all categories above the "no problem" option. They collapsed responses to all CIS items into a dichotomous response of "no problem" (all responses to Option 0) or "at least some problem" (all responses to Options 1–4). We replicated their coding scheme with the data used in the GRM and reanalyzed the data to examine the consequences of collapsing the categories.

Because the items were now dichotomous, the GRM was no longer appropriate. Rather, a two-parameter logistic (2PL) model was employed, which appropriately models dichotomous item responses. The interpretation of a 2PL model is similar to that of the GRM but is parameterized somewhat differently as follows:

$$\Pr(Y_{ij} = 1 | \theta_j) = \frac{\exp\{a_i(\theta_j - b_{ik})\}}{1 + \exp\{a_i(\theta_j - b_{ik})\}}, \theta_j \sim N(0, 1).$$

In the 2PL model, the probability of responding 1, in this case the probability of responding to "at least some problem" for any given CIS item, is conditioned on the latent functional impairment trait of the individual as well as the discrimination and difficulty of the particular item. In the 2PL model, the discrimination parameter still reflects the strength of the relationship of the item with the underlying functional impairment trait. The difficulty parameter now reflects the level of functional impairment needed to have a .50 probability of responding "at least some problem" for any given CIS item.

For both the GRM and 2PL models, item characteristic curves (ICCs) were created for each CIS item to examine the probability of response for each Likert-type scale option as well as the discrimination for various levels of impairment. In addition, the test information function was examined to understand the reliability and amount of information provided by the CIS across varying levels of the functional impairment latent trait. Given the sample size in this study, item-level fit indices, which are sensitive to large samples, were not utilized as a measure of model–data fit. For example, S-χ^2 for all items were significant, $p < .001$, due to the large sample size (Orlando & Thissen, 2000, 2003). Therefore, the following indices guided our analysis of model–data fit as suggested by Kline (2015): the RMSEA, where values less than .06 indicate

Table 1. Descriptive Statistics and CTT Results for the Columbia Impairment Scale.

					Responses by Item Categories[a]							
Item	M	SD	Minimum	Maximum	0	1	2	3	4	α if Deleted	Item–Test Correlation	Item–Rest Correlation
1. Getting into trouble	.39	0.83	0	4	76%	13%	7%	2%	1%	.89	.73	.71
2. Getting along with mother	.47	0.88	0	4	72%	14%	10%	2%	2%	.89	.71	.67
3. Getting along with father	.54	1.00	0	4	71%	14%	9%	3%	3%	.90	.64	.58
4. Feeling unhappy or sad	.51	0.88	0	4	69%	17%	11%	2%	1%	.89	.73	.70
5. Behavior at school	.39	0.85	0	4	78%	11%	7%	2%	2%	.89	.70	.68
6. Having fun	.16	0.54	0	4	90%	6%	3%	1%	1%	.90	.59	.58
7. Getting along with adults other than father/mother	.24	0.65	0	4	85%	8%	5%	1%	1%	.89	.69	.68
8. Feelings of nervousness or worried	.55	0.90	0	4	66%	19%	11%	3%	1%	.90	.61	.56
9. Getting along with brother(s)/sister(s)	.77	1.05	0	4	57%	19%	17%	5%	3%	.89	.67	.61
10. Getting along with other kids his or her age	.34	0.74	0	4	78%	13%	7%	2%	1%	.89	.72	.71
11. Getting involved in activities like sports or hobbies	.37	0.87	0	4	80%	9%	6%	3%	2%	.90	.58	.53
12. With his or her behavior at home	.57	0.92	0	4	65%	19%	12%	3%	2%	.88	.81	.80
13. With his or her schoolwork	.49	0.95	0	4	74%	12%	9%	3%	2%	.89	.68	.64

Note. $N = 69,966$. Data from the 1996–2015 Medical Expenditure Panel Survey Household Component Files. CTT = classical test theory; SD = standard deviation.
[a]Percentages may not total to 100% due to rounding.

acceptable model fit, and the Tucker–Lewis index (TLI) and comparative fit index (CFI), where values greater than .95 indicate acceptable model fit.

All analyses were conducted using R, Version 3.3.3 (R Core Team, 2017). Reliability estimates, item–test and item–rest correlations, and the EFA were all estimated using the Psych package Version 1.8.2 (Revelle, 2017). The IRT models were estimated using the mirt package Version 1.26 (Chalmers, 2012). An R script that recreates the data file and analyses executed in this study is available from the first author upon request.

Results

Descriptive statistics for the 13 CIS items are displayed in Table 1. Responses to the items were overwhelmingly on the low side of the item categories, with most items having 70% or more of responses in the lowest, "no problem" category. The two highest response categories, which would represent the greatest level of functional impairment for each item, were rarely endorsed by respondents. Despite the low variability of responses across the five impairment categories, the CIS demonstrated excellent internal reliability (Cronbach's α = .90). The results of the CTT analysis indicated that overall reliability would not increase by removing any of the items. Item–test correlations indicated that each item of the CIS correlated well with total CIS scores, with correlation values ranging from .58 to .72. The item–rest correlations, which reflect the correlation between an item and the total CIS score with that item removed, ranged from .53 to .80, indicating good discrimination properties for each of the items.

Interitem correlations between all CIS items are displayed in Table 2. The majority of item correlations exceeded a magnitude of .30. Overall, correlations ranged from .27 to .67, indicating that the items were conceptually related to one another but not perfectly collinear. All the correlations fell in the expected positive direction, meaning that none of the items were negatively related to one another. Furthermore, an examination of the Kaiser–Meyer–Olkin measure of sampling adequacy was .93, and Bartlett's test of sphericity was tenable, $\chi^2(78) = 402{,}542.62$; $p < .001$, indicating that the data were suitable for factor analysis.

PAF with no rotation was conducted to identify the number of underlying factors in the CIS data. An examination of the scree plot suggested between a one- and three-factor solutions. While one factor was clearly identifiable in the scree plot, it was possible that up to three factors existed given the second, although smaller, drop in eigenvalues between Factors 3 and 4, with a true leveling off beginning after Factor 3. However, using the Kaiser rule, an examination of eigenvalues supported a one-factor solution, with only the first factor having an Eigenvalue greater than 1.0. An EFA using PAF with a varimax rotation for the one-factor model indicated that all CIS survey items reasonably loaded onto one latent factor, with factor loadings all exceeding .4. The factor loadings are displayed in Table 3. Communalities for all CIS items were at least moderately or strongly related to the single latent factor, with communality values greater than the minimum acceptable value of .20. Finally, model fit indices indicated adequate model fit, RMSEA = .06, $R^2 = .91$. Considering these results, it was determined that the one-factor model was strong enough to assume essential unidimensionality

Table 2. Correlations Among the 13 CIS Items.

CIS Item	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Getting into trouble	—												
2. Getting along with mother	.46	—											
3. Getting along with father	.37	.54	—										
4. Feeling unhappy or sad	.42	.49	.45	—									
5. Behavior at school	.61	.39	.33	.42	—								
6. Having fun	.32	.34	.30	.46	.35	—							
7. Getting along with adults other than father/mother	.51	.45	.37	.44	.49	.47	—						
8. Feelings of nervousness or worried	.32	.33	.31	.52	.30	.38	.37	—					
9. Getting along with brother(s)/sister(s)	.43	.47	.39	.43	.35	.30	.38	.35	—				
10. Getting along with other kids his or her age	.53	.41	.35	.46	.51	.45	.55	.40	.46	—			
11. Getting involved in activities like sports or hobbies	.30	.29	.27	.37	.31	.44	.37	.36	.28	.41	—		
12. With his or her schoolwork	.67	.63	.49	.53	.53	.37	.52	.40	.58	.53	.36	—	
13. With his or her behavior at home	.50	.38	.34	.41	.58	.34	.41	.34	.35	.44	.40	.51	—

Note. N = 69,966. Data from the 1996–2015 Medical Expenditure Panel Survey Household Component Files. CIS = Columbia Impairment Scale.

Table 3. Factor Loadings and IRT Parameters from the GRM and 2PL Models.

Item	Factor Loading	GRM					2PL Model	
		a	b_1	b_2	b_3	b_4	a	b_1
1. Getting into trouble	.71	2.93	0.85	1.42	2.03	2.47	2.81	0.85
2. Getting along with mother	.67	2.54	0.73	1.32	2.10	2.57	2.74	0.70
3. Getting along with father	.56	2.01	0.76	1.37	2.05	2.42	2.30	0.70
4. Feeling unhappy or sad	.69	2.32	0.63	1.32	2.20	2.78	2.33	0.62
5. Behavior at school	.68	2.43	0.96	1.48	2.13	2.58	2.31	0.98
6. Having fun	.57	2.34	1.57	2.12	2.81	3.26	2.34	1.59
7. Getting along with adults other than father/mother	.69	2.93	1.20	1.72	2.39	2.87	2.82	1.21
8. Feelings of nervousness or worried	.55	1.56	0.60	1.49	2.59	3.37	1.55	0.61
9. Getting along with brother(s)/sister(s)	.61	1.93	0.25	0.95	1.94	2.62	1.95	0.24
10. Getting along with other kids his or her age	.71	2.77	0.93	1.55	2.32	2.85	2.84	0.92
11. Getting involved in activities like sports or hobbies	.52	1.56	1.26	1.83	2.56	3.09	1.58	1.25
12. With his or her behavior at home	.81	3.78	0.46	1.10	1.82	2.30	4.49	0.43
13. With his or her schoolwork	.64	2.07	0.83	1.35	2.08	2.61	2.03	0.83
Fit Indices		GRM					2PL Model	
RMSEA		.068					.082	
TLI		.955					.956	
CFI		.970					.963	

Note. N = 69,966. Data from the 1996–2015 Medical Expenditure Panel Survey Household Component Files. 2PL = 2 parameter logistic model; CFI = comparative fit index; GRM = graded response model; IRT = item response theory; RMSEA = root mean square error of approximation; TLI = Tucker–Lewis index.

(Edelen & Reeve, 2007), and the assumption of local independence was also tenable. It was, therefore, appropriate to proceed with IRT analysis.

The discrimination and difficulty parameters for the GRM are displayed in Table 3. The model fit indices for the GRM indicated adequate fit, RMSEA = .069, CFI = .97, TLI = .97. The discrimination estimates ranged from 1.56 to 3.78, generally falling in the middle of this range. Overall, the discrimination estimates indicated that each item has a strong relationship with the latent functional impairment trait measured by the CIS. This is especially true for the items related to problems with behavior at home (a = 3.78), problems with getting into trouble (a = 2.93), and problems getting along with adults (a = 2.93).

The difficulty parameters for all items were above zero. As discussed above, difficulty parameters are placed on a z-score scale, meaning that a person with average functional impairment would have a score of zero on the underlying impairment trait as estimated by the GRM. Accordingly, the estimates for the difficulty parameters indicate that those who are highly impaired are more likely to respond to the high categories of impairment for each item. The results are displayed visually through ICCs in Figure 1. The ICCs indicate that each item discriminates well among high-impairment respondents but provides little to no discrimination for an impairment level below θ = 0. As would be expected, low-impairment respondents were most likely to select Option 0 (no problem), while Option 4 (a very big problem) provided
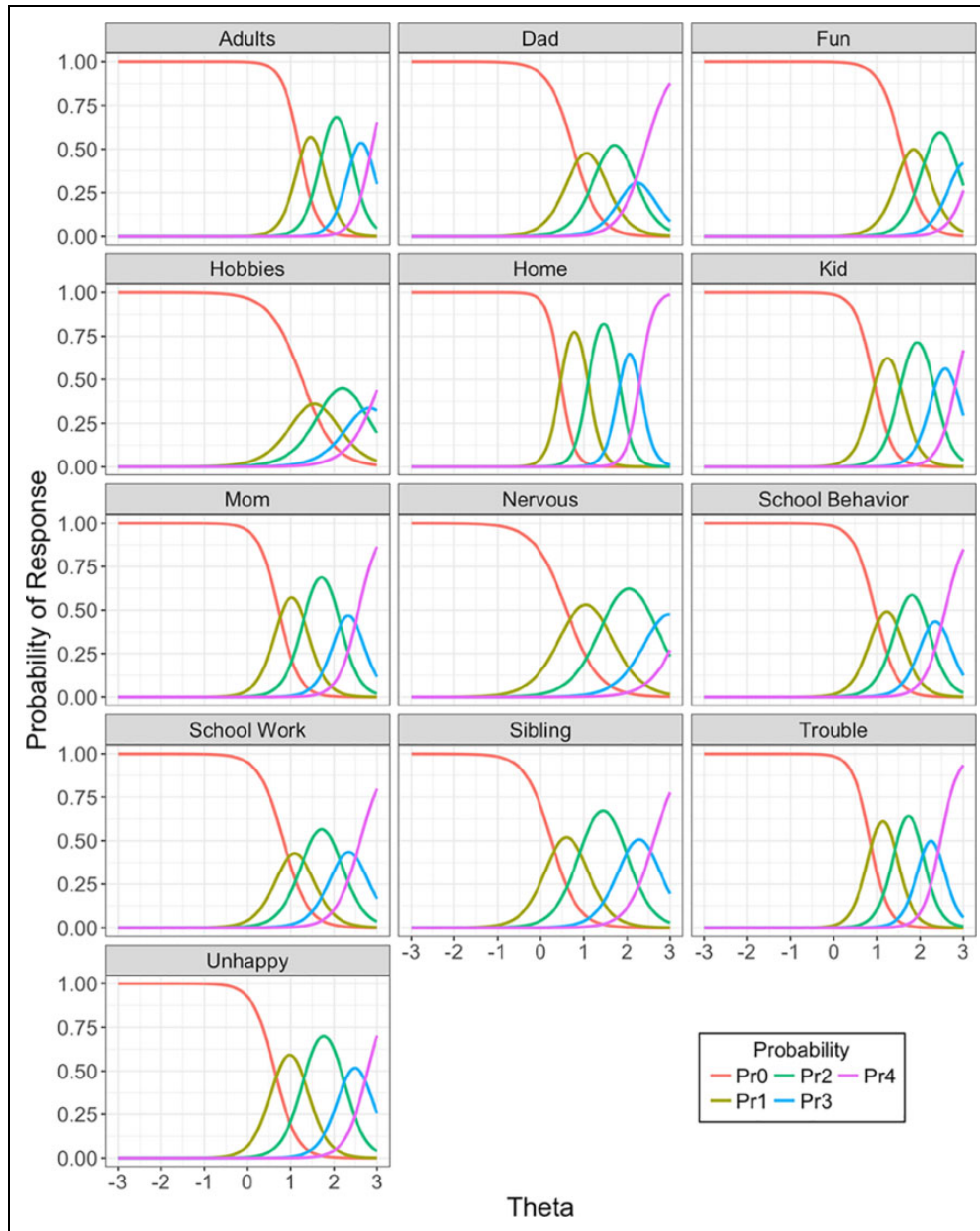
Figure 1. Graded response model item characteristic curves.

high discrimination among respondents with high levels of impairment.

For most items, the probability of respondents selecting Options 2 and 4 are lower than Option 3 and considerably lower than the extreme selection choices of zero or five, indicating that the middle categories of the scale may not be necessary for use in practice. The test information function indicates that the CIS survey reaches the maximum amount of reliability for those with an impairment level of about $\theta = 2$, with the overall greatest amount of information provided for high-impairment respondents, θ higher than zero (Figure 3). However, it is important to note that while reliability was maximized in the

GRM around $\theta = 2$, the overall reliability was respectable across a wide range of the latent functional impairment trait.

The discrimination and difficulty parameters for the 2PL model are displayed in Table 3. Similar to the GRM, the discrimination parameters were all quite high, indicating a strong relationship between each CIS item and the latent functional impairment trait. The items related to problems with behavior at home ($a = 4.49$), problems with getting into trouble ($a = 2.81$), and problems getting along with adults ($a = 2.82$) remained the items with the highest discrimination values. Interestingly, the estimates of the difficulty parameters between the 2PL models were almost identical to the first
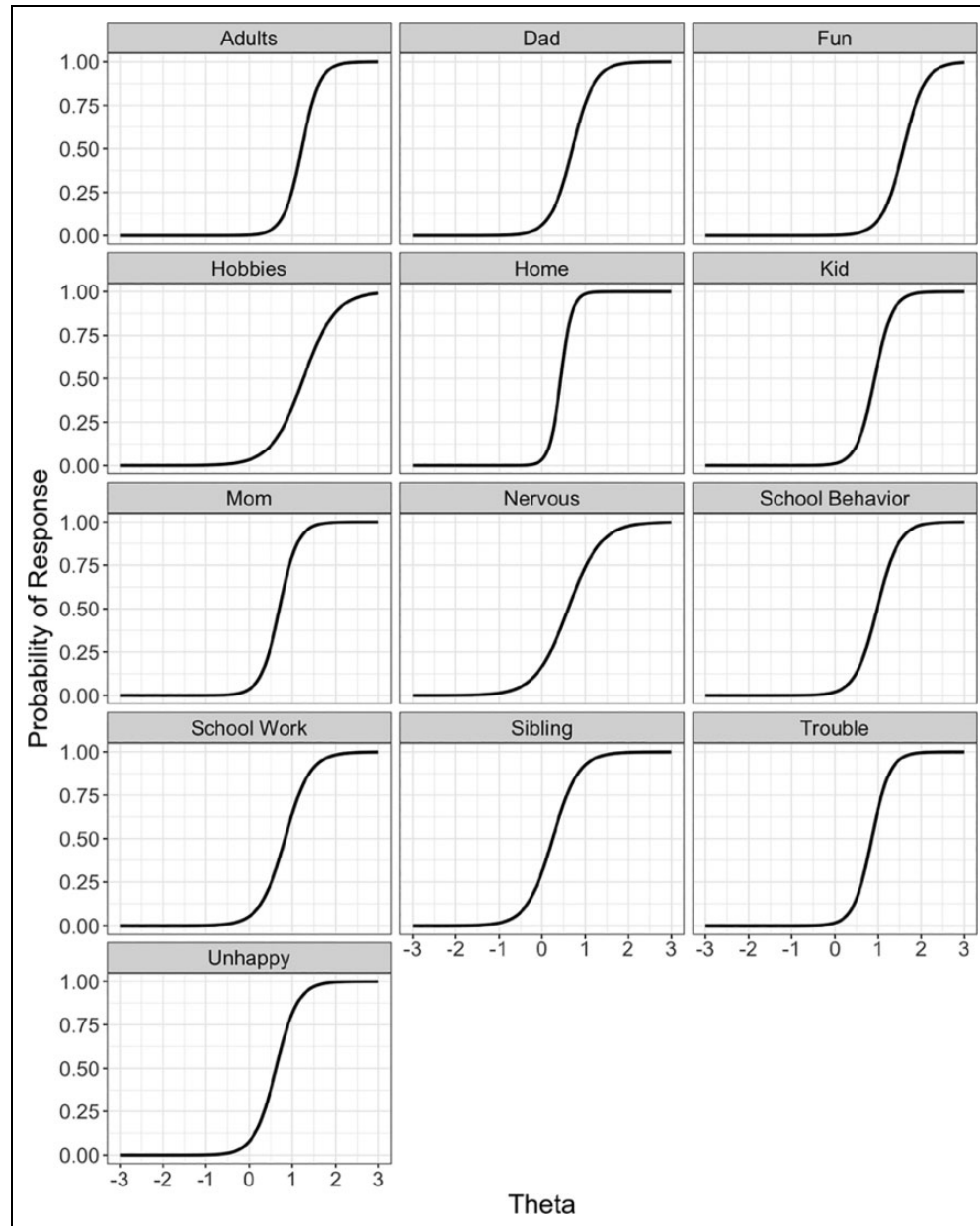
Figure 2. Item characteristic curves for collapsed responses.

threshold estimate of the GRM (see both b_1 columns in Table 3). The model fit indices for the 2PL model also indicated adequate fit, RMSEA = .08, CFI = .96, TLI = .96. ICCs for the 2PL model are displayed in Figure 2. The overall location of the ICCs in the 2PL model mimic the GRM, such that they are all shifted to the right side of the underlying functional impairment trait.

Because we estimated two separate IRT models, we further extended the supplementary analysis to examine whether the values of the impairment trait estimated for each individual in the sample significantly differed between the GRM and 2PL models. To do this, we calibrated the GRM and 2PL model for all respondents and estimated separate θ values under each

model using the expected a posteriori scoring method (Bock & Aitkin, 1981). A correlation coefficient was then calculated to examine whether the θ values significantly differed using the separate IRT models. The results of the analysis indicated that the θ estimates under each IRT model were strongly correlated with one another ($r = .98, p < .001$).

The test information function depicting the reliability of the collapsed CIS is compared to the full five-category CIS in Figure 3. While the reliability of the five-category CIS derived from the GRM is maximized for high-impairment individuals (around $\theta = 2$), the reliability of the two-category CIS derived from the 2PL model is maximized for those closer to average functional impairment (around $\theta = 0.5$). Overall, the amount of
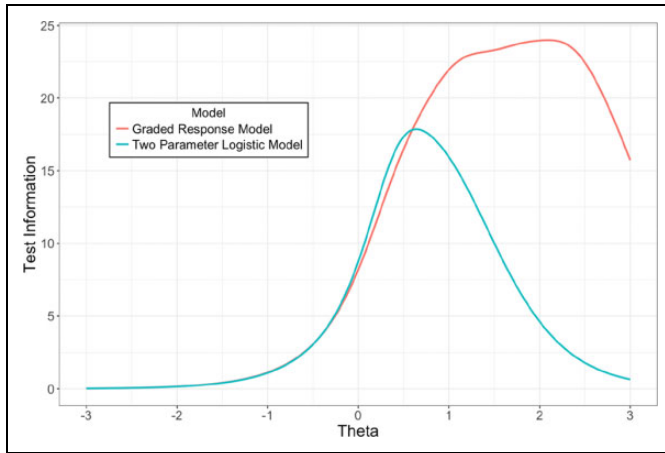
Figure 3. Test information function from the GRM and 2PL Models. GRM = graded response model, 2PL model = two-parameter logistic model.

information provided by the GRM is higher than the 2PL, but the difference is most pronounced for those with above average functional impairment. Indeed, the information function is virtually identical between the two IRT models until around $\theta = 0.5$, at which point the models diverge in amount of information. While the full five-category CIS may be more reliable for higher impairment individuals, the fact that the five-category and two-category IRT models estimate almost identical values for each respondent's latent functional impairment trait provides preliminary evidence that the use of a collapsed, two-category CIS may be reliable for use in research and clinical practice. Nonetheless, in both models, the test information function is respectable across a wide range of the underlying impairment trait. Because the θ estimates are on the *z*-score scale, we would expect an individual with an average level of functional impairment to have a θ score of zero. In both the GRM and 2PL models, the amount of information provided at $\theta = 0$ is around 10. As Embretson and Reise (2000, p. 270) point out, a value of 10 on the test information function equates to a reliability coefficient of .90, which is very reliable. Only on the extreme ends of the impairment trait as estimated by the 2PL model does the CIS exhibit little to no reliability. For the full, five-category CIS, reliability tends to be very low on the −1 to −3 range of the underlying impairment trait.

Discussion and Conclusion

The role of measuring functional impairment holds an important place in research, clinical practice, and service provision for children and adolescents. In addition to federal health policy mandating the measurement of functional impairment for the distribution of mental health block grant funds (Canino, 2016; Ringeisen et al., 2017), behavioral health clinicians rely on assessments of functional impairment to render appropriate diagnoses for mental disorders. Responding to the growing need to track SED among children and adolescents at the local, state, and national level, the CIS was developed in the early

1990s and has remained one of the several popular scales intended to measure functional impairment in children and adolescents. Yet, despite the popularity of the CIS in research and clinical practice, very few studies have specifically examined the psychometric properties to the scale, specifically at the item level. In this study, we addressed the limitations of previous psychometric analyses of the CIS by performing the first and only IRT analysis of the scale using a nationally representative sample of noninstitutionalized Americans.

There are several important conclusions to be taken from the results of the analysis. In relation to previous psychometric analyses of the scale, our findings from the CTT analysis of the CIS provide further support for the reliability of the scale demonstrated by previous findings. In fact, the overall reliability of the CIS in the MEPS sample used in this analysis was greater than the α values reported by all of the community-based studies described above (Olfson et al., 2015; Starin et al., 2014; Yeh et al., 2004). One point of contention between the initial pilot testing of the CIS and subsequent psychometric analyses of the scale relates to the dimensionality of the items. Based on their sample of parents in New York, Bird et al. (1993, 1996) found evidence of a unidimensional, global functional impairment factor, while Singer et al. (2011) found evidence of a three-factor solution in their sample of mothers in Pennsylvania. One possible explanation of the discrepancy between these study findings is that both analyses were performed on relatively small convenience samples. Accordingly, the factor analyses conducted in both studies may be more a reflection of how the CIS functions in those samples rather than the broader population. The analyses performed in this study address the sampling issues of previous studies by exploring the factor structure of the CIS in a more representative population of Americans. Specifically, the results of our analyses lend support to the CIS as essentially unidimensional. Although we did find the presence of potentially three factors, all 13 CIS items loaded strongly onto the first factor, while the second and third factors had relatively minor loadings and eigenvalues less than one for the third factor.

Following our analysis of the factor structure of the CIS, we performed IRT modeling to examine the psychometric properties of the scale across the latent continuum trait of functional impairment. Taken together, the results of the two IRT models provided strong support for the discriminatory properties of each CIS item. Discrimination parameters for both the GRM and 2PL models were very similar, and all of them were high, indicating that each CIS item has a strong underlying relationship with the latent functional impairment trait. In examining both the GRM for the full five-category response option and the 2PL model for the collapsed two-category response option, the results indicated that the CIS most reliably measures functional impairment for children and adolescents who are at average impairment or very functionally impaired. That is, the CIS demonstrated comparatively less, but still very respectable, reliability for those who were low impairment or functioning normally.

These findings have important implications for the health professions. In recent years, federal health policy makers have engaged in discussions about the lack of an existing comprehensive epidemiologic behavioral health surveillance system for children and adolescents. We believe that the CIS would be a viable measure for inclusion in such a system, should it come to fruition. The CIS has a short administration time, does not require administration by a clinician, and based on the results of our analyses demonstrates good psychometric properties. Moreover, its inclusion in the MEPS data would allow national comparisons to be made with the general population. Based on the results of our analysis, we also believe that the CIS would be a desirable instrument for state mental health authorities to utilize in estimating and reporting the prevalence of SED to federal funding agencies, especially given the instrument's strong discrimination properties and respectable reliability across a wide range of the underlying functional impairment trait.

Based on the results of our analyses, there are several areas for future research regarding the psychometric properties of the CIS. Our analysis mostly focused on issues related to the reliability of the CIS at the item and scale level. Although our analysis drew on nationally representative data, rather than community-based samples like those in previous psychometric analyses of the CIS (Singer et al., 2011; Steinhausen & Winkler Metzke, 2001; Zanon, Tomassoni, Gargano, & Granai, 2016; Zielinski, Wood, Renno, Whitham, & Sterling, 2014), we were unable to examine the validity of the CIS against other scales due to their absence in the MEPS data. That is, other scales measuring functional impairment were not administered in the MEPS, and we were therefore unable to examine the validity of the MEPS using more generalizable data. One area of future research would be to examine the construct validity of the CIS with other scales measuring functional impairment in other nationally representative data sets, where possible. It would also be worthwhile for future research to consider applying the same IRT models utilized in our analysis to data collected from community-based clinical samples. Such an analysis would allow for the comparison of discrimination parameters, difficulty parameters, ICCs, and test information functions on the same scale drawn from two populations.

Using the results of our study as a starting point, future researchers could continue to leverage IRT to further examine psychometric properties of the CIS. Several areas of future IRT analyses on the CIS are possible. The first would be to take advantage of the longitudinal nature of the MEPS data to examine the extent of item parameter drift in the CIS. It may be the case that the continued use of the CIS over time has affected accuracy of the CIS item parameters, which may have an impact on the functional impairment latent trait estimated by the model over time (Wells, Subkoviak, & Serlin, 2002). Another opportunity for future research would be to examine differential item functioning among respondents of the CIS. To date, no studies have examined the reliability of the CIS across different groups of individuals. DIF analysis would allow us to understand whether the CIS is biased in measuring impairments across social groups, for example, across gender or race.

Researchers could also further capitalize on the MEPS data to analyze subgroup differences in the factor structure of the CIS. For those individuals exhibiting clinically significant levels of functional impairment, the factor structure of the CIS may be different than individuals with average or low levels of impairment. Finally, future research could also utilize the MEPS data to examine other possibilities in collapsing item responses. We made our decision on which items to collapse based on how the scale was utilized in previous research (see Goudie, Havercamp, Jamieson, & Sahr, 2013), but other decisions surrounding which adjacent categories to collapse may have an impact on the psychometric properties of the scale. If addressed in the future, these studies will add to a small but growing body of evidence that further establishes the reliability and validity of the CIS.

Declaration of Conflicting Interests

Funding

ORCID iD

Brandon K. Attell 🔟 https://orcid.org/0000-0002-4370-3924

References

Agency for Healthcare Research and Quality. (2017). Medical Expenditure Panel Survey. Retrieved from https://meps.ahrq.gov/mepsweb/

Barch, D. M., Gotlib, I. H., Bilder, R. M., Pine, D. S., Smoller, J. W., Brown, C. H., . . . Farber, G. K. (2016). Common measures for national institute of mental health funded research. *Biological Psychiatry*, *79*, e91–e96.

Bird, H. R., Andrews, H., Schwab-Stone, M., Goodman, S., Dulcan, M., Richters, J., . . . Gould, M. S. (1996). Global measures of impairment for epidemiologic and clinical use with children and adolescents. *International Journal of Methods in Psychiatric Research*, *6*, 295–307.

Bird, H. R., Canino, G. J., Davies, M., Zhang, H., Ramirez, R., & Lahey, B. B. (2001). Prevalence and correlates of antisocial behaviors among three ethnic groups. *Journal of Abnormal Child Psychology*, *29*, 465–478.

Bird, H. R., & Gould, M. S. (1995). The use of diagnostic instruments and global measures of functioning in child psychiatry epidemiological studies. In F. C. Verhulst, & H. M. Koot (Eds.), *The epidemiology of child and adolescent psychopathology* (pp. 86–103). New York, NY: Oxford University Press.

Bird, H. R., Shaffer, D., Fisher, P., & Gould, M. S. (1993). The Columbia Impairment Scale (CIS): Pilot findings on a measure of global impairment for children and adolescents. *International Journal of Methods in Psychiatric Research*, *3*, 167–176.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Brannan, A. M., Brashears, F., Gyamfi, P., & Manteuffel, B. (2012). Implementation and development of federally-funded systems of care over time. *American Journal of Community Psychology*, *49*, 467–482.

Brennan, E. M., Nygren, P., Stephens, R. L., & Croskey, A. (2016). Predicting positive education outcomes for emerging adults in mental health systems of care. *The Journal of Behavioral Health Services & Research*, *43*, 564–581.

Burns, B. J., Costello, E. J., Angold, A., Tweed, D., Stangl, D., Farmer, E. M., & Erkanli, A. (1995). Children's mental health service use across service sectors. *Health Affairs*, *14*, 147–159.

Canino, G. (2016). The role of measuring functional impairment in assessing serious emotional disturbance in children. Report commissioned by the National Academies of Sciences, Engineering. Retrieved from https://pdfs.semanticscholar.org/8df5/21baffa7 d720e15a934e589962d96f509f34.pdf.

Canino, G., Costello, E. J., & Angold, A. (1999). Assessing functional impairment and social adaptation for child mental health services research: A review of measures. *Mental Health Services Research*, *1*, 93–108.

Canino, G. J., Fisher, P. W., Alegria, M., & Bird, H. R. (2013). Assessing child impairment in functioning in different contexts: Implications for use of services and the classification of psychiatric disorders. *Open Journal of Medical Psychology*, *2*, 29–34.

Cattell, R. B. (1966). The meaning and strategic use of factor analysis. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 174–243). Chicago, IL: Rand McNally.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29.

Child, D. (2006). *The essentials of factor analysis* (3rd ed.). New York, NY: Continuum International.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*, 5–18.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Ezpeleta, L., Granero, R., Osa, N. D. L., Doménech, J. M., & Bonillo, A. (2006). Assessment of functional impairment in Spanish children. *Applied Psychology*, *55*, 130–143.

Gardner, W., Lucas, A., Kolko, D. J., & Campo, J. V. (2007). Comparison of the PSC-17 and alternative mental health screens in an at-risk primary care sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*, 611–618.

Garland, A. F., Lau, A. S., Yeh, M., McCabe, K. M., Hough, R. L., & Landsverk, J. A. (2005). Racial and ethnic differences in utilization of mental health services among high-risk youths. *American Journal of Psychiatry*, *162*, 1336–1343.

Gold, L. H. (2014). DSM-5 and the assessment of functioning: The World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0). *Journal of the American Academy of Psychiatry and the Law Online*, *42*, 173–181.

Gordon, M., Antshel, K., Faraone, S., Barkley, R., Lewandowski, L., Hudziak, J. J., . . . Cunningham, C. (2006). Symptoms versus impairment: The case for respecting DSM-IV's Criterion D. *Journal of Attention Disorders*, *9*, 465–475.

Goudie, A., Havercamp, S., Jamieson, B., & Sahr, T. (2013). Assessing functional impairment in siblings living with children with disability. *Pediatrics*, *132*, e476–e483.

Gould, M. S., Munfakh, J. L. H., Lubell, K., Kleinman, M., & Parker, S. (2002). Seeking help from the internet during adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*, 1182–1189.

Hammond, S. (2006). Using psychometric tests. In G. M. Breakwell, S. Hammond, C. Fife-Schaw, & J. A. Smith (Eds.), *Research methods in psychology* (3rd ed., pp. 182–209). Thousand Oaks, CA: Sage.

Harris, E. S., Canning, R. D., & Kelleher, K. J. (1996). A comparison of measures of adjustment, symptoms, and impairment among children with chronic medical conditions. *Journal of the American Academy of Child & Adolescent Psychiatry*, *35*, 1025–1032.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, *38*, II28–II42.

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. New York, NY: Routledge; Taylor & Francis.

Holden, E. W., Santiago, R. L., Manteuffel, B. A., Stephens, R. L., Brannan, A. M., Soler, R., . . . Zaro, S. (2003). Systems of care demonstration projects: Innovation, evaluation, and sustainability. In A. J. Pumariega, & N. C. Winters (Eds.), *The handbook of child and adolescent systems of care: The new community psychiatry* (pp. 432–458). San Francisco: Jossey-Bass.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151.

Klein, D. N., Dougherty, L. R., & Olino, T. M. (2005). Toward guidelines for evidence-based assessment of depression in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*, 412–432.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.

Klomek, A. B., Kleinman, M., Altschuler, E., Marrocco, F., Amakawa, L., & Gould, M. S. (2011). High school bullying as a risk for later depression and suicidality. *Suicide and Life-Threatening Behavior*, *41*, 501–516.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Berlin, Germany: Springer Science & Business Media.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163.

Kramer, T. L., Phillips, S. D., Hargis, M. B., Miller, T. L., Burns, B. J., & Robbins, J. M. (2004). Disagreement between parent and

adolescent reports of functional impairment. *Journal of Child Psychology and Psychiatry*, 45, 248–259.

Lambert, M. C., January, S. A. A., Epstein, M. H., Spooner, M., Gebreselassie, T., & Stephens, R. L. (2015). Convergent validity of the behavioral and emotional rating scale for youth in community mental health settings. *Journal of Child and Family Studies*, 24, 3827–3832.

Manteuffel, B., Stephens, R. L., & Santiago, R. (2002). Overview of the national evaluation of the comprehensive community mental health services for children and their families program and summary of current findings. *Children's Services: Social Policy, Research, and Practice*, 5, 3–20.

Manteuffel, B., Stephens, R. L., Sondheimer, D. L., & Fisher, S. K. (2008). Characteristics, service experiences, and outcomes of transition-aged youth in systems of care: Programmatic and policy implications. *The Journal of Behavioral Health Services & Research*, 35, 469–487.

McMahon, R. J., & Frick, P. J. (2005). Evidence-based assessment of conduct problems in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, 34, 477–505.

Merikangas, K. R., Bromet, E. J., & Druss, B. G. (2017). Future surveillance of mental disorders in the United States: Count people, not disorders. *JAMA Psychiatry*, 74, 431–432.

Miller, M., & Hinshaw, S. P. (2010). Does childhood executive function predict adolescent functional outcomes in girls with ADHD? *Journal of Abnormal Child Psychology*, 38, 315–326.

National Academies of Sciences, Engineering, and Medicine. (2016). *Measuring serious emotional disturbance in children: Workshop Summary*. Washington, DC: National Academies Press.

Olfson, M., Druss, B. G., & Marcus, S. C. (2015). Trends in mental health care among children and adolescents. *New England Journal of Medicine*, 372, 2029–2038.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27, 289–298.

Nguyen, T. H., Han, H. R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The Patient: Patient-Centered Outcomes Research*, 7, 23–35.

Painter, K. (2012). Outcomes for youth with severe emotional disturbance: A repeated measures longitudinal study of a wraparound approach of service delivery in systems of care. *Child & Youth Care Forum*, 41, 407–425.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://R-project.org/

Revelle, W. (2017). *Psych: Procedures for Personality and Psychological Research*. Evanston, IL: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psychVersion1.7.8

Ringeisen, H., Stambaugh, L., Bose, J., Casanueva, C., Hedden, S., Avenevoli, S., . . . Copeland, W. E. (2017). Measurement of childhood serious emotional disturbance: State of the science and issues for consideration. *Journal of Emotional and Behavioral Disorders*, 25, 195–210.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded responses. *Psychometrika*, *Monograph Supplement*, 34, 100.

Singer, J. B., Eack, S. M., & Greeno, C. M. (2011). The Columbia Impairment Scale: Factor analysis using a community mental health sample. *Research on Social Work Practice*, 21, 458–468.

Slocum-Gori, S. L., Zumbo, B. D., Michalos, A. C., & Diener, E. (2009). A note on the dimensionality of quality of life scales: An illustration with the Satisfaction with Life Scale (SWLS). *Social Indicators Research*, 92, 489–496.

Snyder, F. J., Roberts, Y. H., Crusto, C. A., Connell, C. M., Griffin, A., Finley, M. K., . . . Kaufman, J. S. (2012). Exposure to traumatic events and the behavioral health of children enrolled in an early childhood system of care. *Journal of Traumatic Stress*, 25, 700–704.

Starin, A. C., Atkins, M. S., Wehrmann, K. C., Mehta, T., Hesson-McInnis, M. S., Marinez-Lora, A., & Mehlinger, R. (2014). Moving science into state child and adolescent mental health systems: Illinois' evidence-informed practice initiative. *Journal of Clinical Child & Adolescent Psychology*, 43, 169–178.

Steinhausen, H. C., & Winkler Metzke, C. (2001). Global measures of impairment in children and adolescents: Results from a Swiss community survey. *Australian & New Zealand Journal of Psychiatry*, 35, 282–286.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. Oxford, England: Oxford University Press, USA.

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, 34, 120–151.

Vishnevsky, T., Strompolis, M., Reeve, C. L., Kilmer, R. P., & Cook, J. R. (2012). Using latent growth curve modeling to examine changes in mental health outcomes for children enrolled in a system of care. *American Journal of Orthopsychiatry*, 82, 121.

Walrath, C., Miech, R., Holden, E. W., Manteuffel, B., Santiago, R., & Leaf, P. (2003). Child functioning in rural and nonrural areas: How does it compare when using the service program site as the level of analysis?. *The Journal of Behavioral Health Services & Research*, 30, 452–461.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26, 77–87.

Winters, N. C., Collett, B. R., & Myers, K. M. (2005). Ten-year review of rating scales, VII: Scales assessing functional impairment. *Journal of the American Academy of Child & Adolescent Psychiatry*, 44, 309–338.

Wu, P., Hoven, C. W., Bird, H. R., Moore, R. E., Cohen, P., Alegria, M., . . . Narrow, W. E. (1999). Depressive and disruptive disorders and mental health service utilization in children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 38, 1081–1090.

Yeh, M., Hough, R. L., McCabe, K., Lau, A., & Garland, A. (2004). Parental beliefs about the causes of child problems:

Exploring racial/ethnic patterns. *Journal of the American Academy of Child & Adolescent Psychiatry*, *43*, 605–612.

Yeh, M., McCabe, K., Hough, R. L., Dupuis, D., & Hazen, A. (2003). Racial/ethnic differences in parental endorsement of barriers to mental health services for youth. *Mental Health Services Research*, *5*, 65–77.

Zanon, A., Tomassoni, R., Gargano, M. L., & Granai, M. G. (2016). Reliability and validity of the Columbia Impairment Scale (CIS) for adolescents: Survey among an Italian sample in Lazio Region. *Epidemiology, Biostatistics and Public Health*, *13*, e116501–e11650-8.

Zielinski, K., Wood, J. J., Renno, P., Whitham, S., & Sterling, L. (2014). Examining the validity of the Columbia impairment scale for assessing level of functioning in youth with autism spectrum disorder. *Journal of Child and Adolescent Psychopharmacology*, *24*, 509–512.