Dimensionality, Hierarchical Structure, Age Generalizability, and Criterion Validity of the GAIN's Behavioral Complexity Scale

Kendon J. Conrad

Karen M. Conrad

Jessica Mazza

Barth B. Riley

Rod Funk

Mark A. Stein

Michael L. Dennis

Kendon J. Conrad, School of Public Health, University of Illinois at Chicago and Chestnut Health Systems, Normal, Illinois; Karen M. Conrad, Program Metrics, LLC, Oak Park, Illinois and School of Public Health, University of Illinois at Chicago; Jessica Mazza, School of Public Health, University of Illinois at Chicago; Barth B. Riley and Rod Funk, Chestnut Health Systems, Normal, Illinois; Mark A. Stein, Department of Psychiatry, University of Illinois at Chicago; Michael L. Dennis, Chestnut Health Systems, Normal, Illinois.

ACKNOWLEDGEMENT: This development of this paper was supported by the Center for Substance Abuse Treatment (CSAT), Substance Abuse and Mental Health Services Administration (SAMHSA) via Westat under contract 270-2003-00006 to Dr. Dennis at Chestnut Health Systems in Bloomington, Illinois using data provided by the following grants and contracts from CSAT (TI-11320, TI-11317, TI-11321, TI-11323, TI-11324, TI-11422, TI-11424, TI-11423, TI-11894, TI-11874, TI-11888, TI-11892, TI-11871, TI-13309, TI-13356, TI-13305, TI-13340, TI-13344, TI-13322, TI-13323, TI-13345, TI-13308, TI-13354, TI-13313, TI-14254, TI-14376, TI-14311, TI-14196, TI-14214, TI-14261, TI-14090, TI-14189, TI-14252, TI-14283, TI-14355, TI-14272, TI-14103, TI-14267, TI-14315, TI-14188, TI-14271, TI-15686, TI-15671, TI-15486, TI-15545, TI-15672, TI-15475, TI-15678, TI-15447, TI-15461, TI-15433, TI-15481, TI-15514, TI-15478, TI-15413, TI-15483, TI-15670, TI-15674, TI-15479, TI-15682, TI-15467, TI-15511, TI-15562, TI-13601, TI-13190, TI-12541, TI-00567; Contract 207-98-7047, Contract 277-00-6500), the National Institute on Alcohol Abuse and Alcoholism (NIAAA) (R01 AA 10368), the National Institute on Drug Abuse (NIDA) (R37 DA11323; R01 DA 018183), the Illinois Criminal Justice Information Authority (95-DB-VX-0017), and the Illinois Office of Alcoholism and Substance Abuse (PI 00567). The opinions are those of the authors and do not reflect official positions of the contributing project directors or government.

Correspondence concerning this article should be sent to: Kendon J. Conrad, Senior Research Scientist, Chestnut Health Systems and Professor Emeritus, Health Policy and Administration, University of Illinois at Chicago, School of Public Health (M/C 923), 1603 West Taylor Street Chicago, IL 60612-4310. Phone: 312.996.3185. Email: kjconrad@uic.edu

Abstract

This study used Rasch measurement model criteria and traditional psychometric strategies to examine key psychometric properties of the Behavioral Complexity Scale (BCS), a widely used measure of externalizing disorders that focuses on attention deficit, hyperactivity and conduct disorders. Using a sample of 7,435 persons being screened for substance use disorders, the BCS was found to: (1) be unidimensional, (2) have a hierarchical severity structure, (3) be generalizable to both youth and adults, (4) and meet hypothesized correlations with criterion variables. The *BCS* performed well as a unidimensional measure. The Rasch severity hierarchy of attention deficit to hyperactivity to conduct disorders provided a perspective that suggested that a dimensional measure could be used as an alternative and, in some ways, as an improvement, to categorical diagnosis and common dimensional approaches. The finding of three low severity conduct disorder items also supported a revision of categorical criteria, especially in substance use disorders.

Keywords: externalizing disorders, inattentive disorders, hyperactive/impulsive, conduct disorders, Rasch model, Global Appraisal of Individual Needs

Dimensionality, Hierarchical Structure, Criterion Validity, and Age Generalizability of the GAIN's Behavioral Complexity Scale

The general purpose of this study was to examine the validity of interpretations of the Behavioral Complexity Scale or BCS (Dennis, Chan & Funk, 2006), a measure of externalizing disorders (full description below and in Table 1) that focuses on attention deficit, hyperactivity and conduct disorders. The BCS is part of the Global Appraisal of Individual Needs (GAIN, Chestnut Health Systems, 2010) and, by the end of 2011, was being used mostly in substance use disorders (SUD) treatment settings by over 1700 agencies in 48 states, 6 provinces of Canada, and a half dozen other countries.

Background

Studies in the general population (Achenbach & Edelbrock, 1978; Krueger, 1999; Vollebergh, Iedema, Bijl, deGraaf, Smit, & Ormel, 2001) and in treatment-based samples (Chan, Dennis, & Funk, 2008; Dennis, et al., 2006; Dennis, Dawud-Noursi, Muck, & McDermeit, 2003) demonstrate the existence of three primary dimensions along which the symptoms of the more common mental disorders vary: (a) internalizing disorders (e.g., symptoms of depression, anxiety, somatic disorder, traumatic distress, suicide) characterized by prevalence increasing with age, doubling between pre-adolescence and early 20s; (b) externalizing disorders (e.g., symptoms of attention deficit, hyperactivity, and conduct disorders) characterized by on-set before adolescence with prevalence cut in half by early 20s; and (c) substance use disorders (e.g., symptoms of abuse, dependence, other substance-induced health or psychiatric problems) characterized by on-set during adolescence and peaking in early to mid 20s. The second

dimension, externalizing disorders, is the least studied of the three dimensions, particularly among adults.

The GAIN's BCS scale focuses on externalizing disorders in general and specifically on symptoms related to attention-deficit/hyperactivity disorder (ADHD) and conduct disorder (CD). According to DSM-IV (American Psychiatric Association [APA], 2000), ADHD is comprised of two clusters of symptoms, inattentive behaviors and hyperactive-impulsive behaviors. An individual with ADHD can be diagnosed with ADHD-predominantly **inattentive type**, ADHD-predominantly **hyperactive-impulsive type**, or ADHD-**combined type** (both clusters of symptoms). DSM-IV does not specify which of the disorders, inattentive or hyperactive, is more severe, and prevalence ranges from 6-9% for youth, and approximately 5% for adults in the general population (Wilens, 2004) and from 45-48% for youth and 31-39% for adults in substance abuse treatment (Chan et al., 2008).

Conduct Disorder (CD) refers to a set of more severe disruptive behavioral problems that can include behavioral aggression against others, destruction of property, and serious violations of normative rules, such as persistent truancy and running away from home (APA, 2000). According to DSM-IV, the severity of CD can be specified as **mild** if there are few problems with little harm to others (e.g., lying, truancy, staying out late), **moderate** if there are some more or more moderate symptoms (e.g., stealing without confronting a victim, vandalism), or **severe** if many more or severe symptoms (e.g., forced sex, physical cruelty, use of a weapon, stealing while confronting a victim, breaking and entering). DSM (APA, 2000, p. 96) also explicitly recognizes a high rate of comorbidity between ADHD and CD. The prevalence of CD is approximately 11% among youth and 8 to 10% among adults in the community (Nock, Kazdin,

Hiripi, & Kessler, 2006), and 56-59% among youth and 25-40% among adults in substance abuse treatment (Chan et al., 2008).

While categorical models such as DSM-IV have traditionally placed individuals in distinct diagnoses and subtypes such as those described above, there is growing evidence that externalizing disorders are a result of common, underlying core psychopathological processes as opposed to discrete disorders (Krueger & South, 2009). Because CD only requires 3 of 15 symptoms, even minor changes in the diagnostic criteria have been shown to result in major differences in prevalence (Boyle et al. 1996; Loeber, Burke, Lahey, Winters, & Zera, 2000) and has led many to suggest that symptom severity may be better accounted for as a single dimension or spectrum (Krueger, 1999; Krueger, Markon, Patrick, & Iacono, 2005; Nock et al. 2006; Stein & O'Donnell, 1985) especially in longitudinal studies (such as Barkley, Fischer, Smallish, & Fletcher, 2004) when subjects may display more serious symptoms over time or exhibit changes in the manifestations of an underlying pathology as they get older. Indeed, in their recent meta-analysis, Markon, Chmielewski and Miller (2011) found that shifting from categorical to dimensional measures of severity increased reliability by an average of 15% and validity by an average of 37%.

Unidimensionality

The idea of a spectrum connotes existence of a single construct, i.e., unidimensionality of externalizing disorders. Especially in the measurement context, it also implies a hierarchical structure (grading) of symptom severity. While there is some controversy over the complex relationships of ADHD and CD symptoms, e.g., whether CD is simply a more severe form of ADHD, whether CD develops as children with ADHD get older, whether there are both separate diagnoses of ADHD and CD as well as a hybrid of the two (Szatmari, Boyle, & Offord, 1989), it

is increasingly common to subsume these disorders under the unidimensional construct of externalizing disorders (Cohen, Gotlieb, Kershner, & Wehrspann, 1985; Rapport, LaFond, & Sivo, 2009).

Hierarchical Structure

The use of quantitative psychological models, such as item response theory (IRT) and latent-variable models, has allowed conceptualization of psychopathology in ways that are more continuous, rather than categorical (Krueger & Markon, 2008), and they possess the capability of examining the items and constructs in a severity hierarchy using the proportion/probability of item endorsement as a severity estimator. While CD is often considered to be more severe than inattentive and hyperactive disorders, there was little literature to suggest a hierarchy of severity for inattentive vs. hyperactive subtypes of ADHD. Despite the high prevalence of these disorders, there has been little use of modern psychometric models in research on externalizing disorders as a spectrum, e.g., to examine the validity of grading the severity of symptoms and correspondingly of persons along a spectrum, in contrast to counting equally weighted symptoms in order to classify individuals. Instead, studies have focused on the distinct constructs of inattentiveness, hyperactivity, and conduct/oppositional defiant disorders separately (Conners, Erhardt, & Sparrow, 1999; Epstein & Kollins, 2006; Erhardt, Epstein, Conners, Parker, & Sitarenios, 1999; Frick, 2006).

Modern measurement techniques based in item response theory (IRT, Embretson & Reise, 2000), such as Rasch measurement models (Rasch, 1960; Wright & Stone, 1979) hold promise in addressing these issues since these models can generate item severity calibrations based on the probability of item endorsement. For an individual, a probabilistic estimate of a person's severity level is obtained based on their endorsement of items whose severity has been

estimated as a unidimensional measure. If items misfit, this may suggest they should be revised or dropped. Additionally, examination of the severity hierarchy by demographic characteristics enables us to estimate whether persons in different groups, e.g., gender or age groups, respond in systematically different ways, perhaps indicating bias, even though they are at the same level on the construct.

Item Invariance by Age

Due to changes in externalizing symptoms with development, questions remain about the validity of measures and criteria across age groups, especially for youth vs. adults in regard to ADHD and CD. For example, Buitelaar (2007) suggested from clinical and biological studies that the criteria for diagnosis of ADHD should be rephrased and modified for adults, and that the diagnostic algorithm be modified to require only 4 out of 9 symptoms for adults, and that the age of onset requirement should be increased to 16 years. Furthermore, additional research is needed to determine the validity of subtypes in adults. To help inform these issues, modern measurement provides one method of examining whether the symptom hierarchy is the same for youth compared to adults. If so, this would suggest that a common measure could be used for all age groups. If not, then separate age-group-specific measures or age-adjusted measures may be needed (e.g., Conrad, Dennis, Bezruczko, Funk, & Riley, 2007).

The literature on measurement invariance in externalizing disorders is sparse. Guttmannova, Szanyi, & Cali, (2008) examined the Behavior Problem Index (Peterson & Zill, 1986), which is comprised of 28 parent-reported items of both internalizing and externalizing disorders derived from the Child Behavior Checklist. The measurement invariance of the Behavior Problem Index was supported in terms of factor loadings and thresholds across ethnic groups at each time point and within each ethnic group over time. However, Mezzacappa (2007)

applied item response theory to parent ratings of aggression and delinquency from the Child Behavior Check List (CBCL) and found differences in rates of endorsement due to non-equivalencies in response thresholds across the three different ethnic-racial groups studied: namely, Caucasians, African-Americans, and Hispanics. Non-equivalencies in item response thresholds were identified across all group comparisons, indicating the absence of measurement equivalence for CBCL-defined aggression and delinquency across ethnic-racial groups in this sample. These authors emphasized that the invariance of a measure across demographic groups is key to understanding its validity, e.g., generalizability, and usefulness in various groups and settings. Of course, item invariance by age group, i.e., differential item functioning, is a key issue as well, but we could find no psychometric evidence on this issue for an externalizing disorders spectrum.

Criterion Validity

Another important validity test of the BCS is criterion validity, i.e., evidence based on relations with other theoretically related variables. While this is not a Rasch-based indicator, it is an important traditional validity category (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Testing of criterion validity would typically include examining hypothesized correlations (Campbell & Fiske, 1959) of theoretically grounded criterion variables. Substance abuse would be one key hypothesized criterion since an increased prevalence of substance use disorders (SUD) in youth and adults diagnosed with ADHD and CD has been well established (Achenbach, 1991; Krueger & South, 2009; Sher & Trull, 1994). Of note, individuals with co-morbid ADHD and CD are at an even more elevated risk for substance use and abuse (see Barkley et al., 2004; Biederman, Newcorn, & Sprich 1998; Disney, Elkins, McGue, & Iacono,

1999; Edwards & Kendler, 2012; Wilens, 2004; Wilens, Biederman, & Spencer, 2002) and are at higher risk of other psychiatric disorders and impairments such as legal problems and premature mortality (Kim-Cohen et al., 2003; Simonoff, et al., 2004). Indeed, as Lilienfeld (2003) noted in regard to co-morbidity in general, “individuals who exhibit antisocial behavior are often confronted with adverse life consequences” (p. 288).

Objectives

Therefore, the goal of testing the validity of the BCS was addressed using a combination of Rasch model and traditional measure validation techniques to assess the following hypothesized patterns: (1) dimensionality, whereby a unidimensional measure of externalizing disorders was hypothesized to exist; (2) hierarchical structure, whereby inattentive symptoms and hyperactive/impulsive symptoms would be equal in severity and conduct disorder symptoms would be higher; (3) invariance of the symptom severity hierarchy for age (youth <18 vs. adults); (4) criterion validity interpretations of the BCS in terms of severity based on correlation with other theoretically related variables, i.e., substance problems, HIV risk, emotional problems, and recovery environmental risk.

Methods

Data Source

Data on the 7,435 cases reported in this paper were pooled from 77 substance abuse treatment studies funded by the Center for Substance Abuse Treatment, National Institute on Alcohol Abuse and Alcoholism, National Institute on Drug Abuse, Robert Wood Johnson Foundation and Interventions Foundation. The studies were conducted in a variety of institutional settings, including across adolescent and adult levels of care. All data were collected and managed by the GAIN Coordinating Center located at Chestnut Health Systems in

Normal, Illinois. Founded in 1973, Chestnut Health Systems is a large non-profit behavioral health firm that focuses on providing adult and adolescent mental health and addiction treatment services, employee and student assistance programs, disease management services, justice programs, prevention, and applied research on how to make the services more effective. Interviews were conducted during intake/screening using the GAIN (described below) as part of clinical practice or specific research studies under their respective voluntary consent procedures with identifiers subsequently encrypted before making the data available for analysis under the supervision of Chestnut Health System's Institutional Review Boards. Research studies were conducted with general consents under federal guidelines (42 CFR Part 2) that explicitly allow record abstraction for the purpose of program evaluation and development as long as the data are de-identified and kept confidential. Data obtained since the implementation of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) were covered by formal data sharing agreements between Chestnut Health Systems and each of the participating agencies per 45 CFR Parts 160 and 164, Subparts A and E. All interviews were conducted by interviewers with three to four days of training followed by rigorous field-based certification procedures. Field interviewers had ongoing supervision by local trainers who were trained and certified by Chestnut staff on the use of the GAIN. Full details about the GAIN in general and BCS specifically may be obtained at Dennis (n. d.) GAIN home page: http://www.chestnut.org/LI/gain/

Background Characteristics of the Sample

The sample was predominantly under 18 years of age (73%) and male (67%). Almost half was Caucasian (45%), a quarter was African American (26%), and the remainder Hispanic (11%) or mixed/other (18%). Of the top five most severe primary drugs reported, marijuana was

reported by 49% of the sample. The drug least often reported was opioids at 5%. Other drugs reported included amphetamines (11%), cocaine (11%), and alcohol (20.5%). Almost 3% percent of the sample reported other drugs. Using DSM-IV categorical criteria described below, 59% had an externalizing disorder, including 43% with ADHD and 51% with CD (8% had ADHD disorder only, 16% had CD disorder only, 35% had both ADHD and CD disorders).

Description of the Measure

Administered by an interviewer, the BCS asks clients whether they have had two or more problems in the past year related to the 33 symptoms of ADHD and CD that are given in lay terms but based on *DSM-IV* criteria (American Psychiatric Association, 2000). This basis in DSM-IV supports the content validity of the BCS (Messick, 1989). The item stem reads: *"During the past 12 months, have you done the following things two or more times?"* The response format is Yes/No (coded: no=0, yes=1).

BCS Subscales and DSM-IV Criteria

The three BCS subscales are the Inattentive Disorder Scale (IDS; 9 items), Hyperactivity-Impulsivity Scale (HIS; 9 items), and Conduct Disorder Scale (CDS; 15 items). Responses on each of the three subscales (IDS, HIS, CDS) of the BCS are used to suggest clinical diagnoses. If 6 or more items are endorsed on the Inattentive Disorder Scale (IDS), it is suggested that a diagnosis for ADHD – Inattentive type be considered. If 6 or more items are endorsed on the Hyperactivity Impulsivity Scale (HIS), it is suggested that a diagnosis for ADHD – Hyperactive type is considered. If 6 items on both IDS and HIS are endorsed, the diagnosis of Inattentive/Hyperactive Combined Type is considered. On the Conduct Disorder Scale (CDS), 3 or more items must be endorsed to suggest a diagnosis. The subscale names, item stems, and item labels are shown in Table 1.

Analysis Procedures

Objective 1. Estimate dimensionality whereby a unidimensional measure of externalizing disorders was hypothesized to exist. As described by Embretson & Reise (2000) and others, the Rasch model (Rasch, 1960) is the item response theory model that meets requirements, including unidimensionality, of linear, interval measurement (Wright & Stone, 1979; Bond & Fox, 2007). Our choice of the Rasch model was also supported by its generation of fit statistics that enable setting stable criteria for determining whether individual items contribute to or detract from unidimensionality.

The BCS was analyzed with a Rasch dichotomous model (Rasch, 1960; Wright & Stone, 1979) with *Winsteps version 3.72.0* statistical software (Linacre, 2011) to obtain linear, interval measures. The dichotomous model estimates the probability that a respondent will choose a particular response category for an item as:

$$\ln \frac{Pni}{1-Pni} = B_n - D_i,$$

where ln is the natural logarithm, P_{ni} is the probability of respondent n endorsing item i, $1\text{-}P_{ni}$ is the probability of respondent n not endorsing item i, B_n is the ability/severity of respondent n, D_i is the difficulty of item i.

Dimensionality. Analysis of the dimensionality of the BCS involved a two-step process. First, the measurement dimension of the BCS was estimated using the Rasch model. The variance estimate associated with this measurement dimension was obtained from the item-response data by computing standardized residuals, i.e., (observed - expected)/(model standard error). Second, a principal component analysis of the standardized residuals was used to determine whether substantial sub-dimensions existed within the BCS items (Linacre, 1998a,b; Smith, 2002). If the items measure a single latent dimension as estimated by the Rasch model,

then the remaining residual variance should reflect random variation. As suggested by Embretson & Reise, 2000, we chose a ratio of variance of at least 3 to 1 in the first principal component compared to the variance of the first component of residuals. We also tested dimensionality using Linacre's (1998b) procedure as follows. We extracted two subsets of items representing the opposite poles of the construct, i.e., less severe vs. more severe, and measured each subject on each subset of items. We cross-plotted the subject measures to check whether the plots were on the diagonal and obtained correlation coefficients. We also cross-plotted original item calibrations against the new item calibrations for the separate subset scales. These lines should shadow each other except for scaling differences. Additional criteria for unidimensionality were employed using item fit statistics as discussed next.

Fit statistics as dimensionality criteria for items. Rasch analysis provides fit statistics to test assumptions of fundamental measurement for items (Wright & Stone, 1979), a structural aspect of validity (Messick, 1995). The Rasch model provides two indicators of misfit: infit and outfit. The infit is sensitive to unexpected, e.g., too much random behavior, affecting responses to items near the person ability level or item difficulty level and the outfit is outlier sensitive. For examples, please Linacre & Wright (1994) http://www.rasch.org/rmt/rmt82a.htm. Understanding poor fit can lead to dropping or improving items because they fail to perform in a manner consistent with the principal measurement dimension, i.e., detract from unidimensional measurement.

Mean square statistics are defined such that the model-specified uniform value of randomness is 1.0 (Wright & Stone, 1979). The items with high outfit mean squares are items with more unexpected responses than are consistent with the model at the tails (e.g., endorsing high severity items but not low severity items). Using Wilson's (2005) criterion of >1.33 and

<0.75, an item was regarded as misfitting if its mean squares on both infit and outfit were higher than 1.33 or lower than 0.75, i.e., the latter being over-fit.

Person reliability and alpha. Reliability is a structural aspect of construct validity (Messick, 1989). Reliability was estimated with Cronbach's alpha and Rasch person reliability statistics. Both indices reflect the proportion of variance of the person scores or measures to total variance (i.e., including measurement error). Unlike Cronbach's alpha, Rasch person reliability is based on the estimated locations of persons along the measurement continuum, excluding those with measures reflecting extreme (zero or perfect) scores and including cases with missing data. For both indices, our criterion for acceptability was .80.

For a complete treatment of Rasch analysis, we recommend Bond & Fox (2007) which includes a glossary of Rasch measurement terminology and Conrad & Smith (2004) for a brief summary with useful references. Terminology may also be accessed online via Rasch Measurement Transactions located at http://www.rasch.org/rmt/.

Objective 2. Test the hierarchical structure, whereby inattentive symptoms and hyperactive/impulsive symptoms would be equal in severity and conduct disorders symptoms would be more severe (higher). The general logic is that, when items are less severe, they are closer to the population norm where most people are located, i.e., having no symptoms or a few mild ones that are relatively common. Also, the less severe items cause less discomfort to others and are more tolerable. As a result, less severe symptoms are likely to occur more frequently. More severe symptoms are farther from the norm, i.e., occur less frequently. They affect others more and therefore have social taboos and/or consequences such as interventions and punishments that are undesirable. Thus, based on the literature, the inattentiveness subscale (IDS) and the hyperactivity/impulsivity subscale (HIS) were

hypothesized to be equivalent in severity while the conduct disorder subscale (CD) was presumably more severe. Rasch analysis provides this item severity hierarchy using the endorsement probabilities of persons and items as described in the equation above and displayed on a Wright map (Wilson, 2005). For more thorough discussion of the interpretation of Wright maps and their role in construct validation, please see http://www.rasch.org/mra/mra-01-10.htm (Lunz, 2010), and http://www.rasch.org/rmt/rmt221.pdf (Baghaei, 2008).

As noted above, the Rasch hierarchy is based on frequency of endorsement which has been found useful in education as well as in hundreds of studies in behavioral health (Conrad & Smith, 2004), and it addresses the substantive or theoretical aspect of construct validity (Messick, 1995). Of course, there are exceptions or variations that can be examined using methods such as differential item functioning analysis (discussed under Objective 3).

Objective 3. Test the Invariance of the Symptom Severity Hierarchy for Age. As Bond and Fox (2007) noted, the Rasch model requires that relative item estimates, i.e., item difficulty estimates, remain invariant across subgroups of persons when members of the groups are at the same level on the trait, e.g., females and males. Examination of differential item functioning (DIF) allows us to test whether items reflect significantly different levels of symptom severity for different groups, i.e., differing item calibrations. This analysis examines the validity issue of construct irrelevance for groups (Messick, 1995).

Bond and Fox suggested that items that show DIF should be investigated to determine what may be inferred about the underlying construct and what DIF implies about the subsamples of persons. In other words, DIF analysis addresses an important construct validity criterion concerning the comparability and fairness of items, their calibrations, and the interpretation of the resulting scores (Messick, 1995). Since the large sample made most DIF contrasts

statistically significant, we chose a clinically significant DIF contrast that was based on $\geq$ .7 logit difference for all comparisons which is approximately half of a standard deviation (SD = 1.35) for the persons. Standards for what is considered an important DIF effect size vary from about .4 to .6 logits (see Longford, Holland, Thayer; 1993; Paek, 2002; Draba, 1977; Elder, McNamara, & Congdon, 2003; Scheunemann & Subhiyah, 1998; Wang, 2000). In this paper, we used the criterion of .7 logit or larger since we believed that most would agree that this is a large, important DIF contrast, and half a standard deviation is a common criterion for clinical significance (Cohen, 1988; Norman, Sloan, & Wyrwich, 2003; Wolf, 1986). The Winsteps procedure for DIF analysis did not include the identification and linking of common items. Rather, a common person measure based on data from both groups (excluding the targeted item) was computed. Persons were then anchored on this common person measure when computing group-specific item calibrations, thereby placing these calibrations on a common metric based on the total sample.

If the items of the BCS were not invariant across subgroups, then changes to the measures would be considered, e.g., dropping items and developing new ones or developing separate measures for certain subgroups. Findings of differential item functioning might also have theoretical implications for the treatment of ADHD, CD, and SUD depending on the subgroup.

Objective 4. Test criterion validity interpretations of the BCS in terms of severity based on correlation with other theoretically related variables. To examine the correlates of externalizing disorders, with other variables as criterion validity criteria (Campbell & Fiske, 1959; Messick, 1989), we set up a pattern of expected correlations, where high >.5 and moderate >.3, using measures that were available in the GAIN. High correlations were expected with

variables derived from constructs that were theoretically closely related to externalizing disorders, i.e., the Substance Problems Scale and Emotional Problem Scale, which includes internalizing and externalizing behaviors that frequently co-occur in individuals with disruptive behavior disorders (see Biederman, Newcorn, & Sprich, 1991; Hinshaw & Nigg, 1999; Jensen, Martin, & Cantwell, 1997). In contrast, only moderate correlations were expected with constructs that are related to externalizing disorders (ExDx) but that are less similar. In this case, we used two variables that should logically be associated with sequelae of inattentiveness, impulsiveness and conduct disorders, i.e., Recovery Environment Risk Index and HIV Risk Scale. In other words, we hypothesized moderate associations based on increased risk taking behavior in individuals being screened for substance use disorders who exhibit externalizing disorders. In addition to estimating the correlations with the full BCS measure, we bar-graphed the mean values of each validation scale for persons within the BCS cut points where 0 logits =no ExDx (Rasch BCS measure below -2), 1=low ExDx (Rasch BCS measure of -1 to -2), 2=moderate ExDx (Rasch measure of -1 to 0), and 3=high ExDx (Rasch measure greater than 0). These should display monotonically increasing average scores on the criterion measures with each increasing BCS severity level. We emphasize that these BCS cut-points were not the result of rigorous evaluation and were set for the purposes of illustration. However, we note that they corresponded roughly to the increasing severity of symptoms, which was a result of objective 2.

Measures used to test criterion validity.

Substance Problems Scale-past year (alpha=.90). The GAIN's Substance Problems Scale is a count of past-year, yes/no symptoms of substance abuse, dependence, or substance induced disorders and is based on DSM-IV (Conrad et al., 2007; Modisette, Hunter, Ives, Funk, & Dennis, 2009).

The HIV Risk Scale (alpha=.86). This is a count of 35 yes/no items related to needle use activities, sexual risk behaviors, and victimization (Conrad, Conrad, Dennis, Riley, & Funk, 2009).

Emotional Problems Scale (alpha=.80). This is an average of items (divided by their range) for recency of mental health problems, memory problems, and behavioral problems and the days (during the past 90 days) of being bothered by mental problems (Modisette et al., 2009).

Recovery Environmental Risk Index (retest Spearman Rho=.75). This is an average of items (divided by their range) for the days (during the past 90 days) of alcohol in the home, drug use in the home, fighting, victimization, being homeless, and structured activities that involved substance use and the inverse percentage of days going to self-help meetings, and involvement in structured substance-free activities (Lennox, Dennis, Scott, & Funk, 2006).

Results

Objective 1. Estimate dimensionality

Dimensionality. The variance explained by the full BCS measure was 37.9%. Only 9.1% of the variance was explained by the first factor of residuals. Therefore, the ratio of the first principal component to the first factor of residuals was greater than 4 to 1 whereas the criterion was only 3 to 1. We interpreted the high variance explained by the principal measurement dimension and the low variance explained by the first factor of residuals as supportive of unidimensionality.

Using Linacre's (1998b) procedure, the two resulting constructs were ADHD vs. CD. When separate ADHD (alpha=.93) and CD (alpha=.85) measures were obtained on clients and the measures were subsequently correlated (n=7,345), the correlation was .76, and the correlation corrected for attenuation due to measurement error was .96. The cross-plots of the two measures

formed a clear diagonal. This was also supportive of unidimensionality. As further support, plotting the original item calibrations against the separate high/low calibrations (actually the ADHD items vs. the CD items) indicated that the recalibrated items shadowed the original calibrations very closely.

Item fit. There were no substantial infit problems (Table 1, Column 4). *SkipSchool/Work* had the highest outfit and infit, but the infit was less than 1.33 so it did not meet the criterion for a substantial problem. Therefore, the fit analysis results were also supportive of the interpretation that the BCS was unidimensional, with the possible exception of *SkipSchool/Work*. Issues with this item were also noted in the DIF analysis discussed below.

Person reliability. Rasch person reliability of BCS scores was strong at .87. Cronbach's alpha was .94.

Objective 2. Test the Hierarchical Structure

Examination of the Wright map (Figure 1) indicates a hierarchy of severity with inattentive disorders at the low, most frequent level; hyperactive/impulsive symptoms at the moderate level; and conduct disorders at the high, least frequent level. However, it is notable that three CD items, i.e., *SkipSchool/Work, LiedConned,* and *StayOut2Late*, were exceptional since they belong conceptually with the most severe construct but were located empirically at the lower level along with the least severe inattentive symptoms.

Objective 3. Test the Invariance of the Symptom Severity Hierarchy for Age

In Figure 2, there were some large (>.7 logit) differences between youth and adults. Specifically, four items, i.e., *NotFollowInstruct, SkipSchool/Work, DestroydPrprty,* and *SetFires,* were easier for youth to endorse. Concurrently, there were five items that were > .7 logit difference that were easier for adults to endorse, i.e., *AbsentMinded, Restless, FeltOnTheGo,*

Wait, and ForcedSex. Specifically, youth tended to endorse serious CD symptoms of vandalism and arson as well as less serious school-related items, *SkipSchool/Work,* a conduct disorder item, and *NotFollowInstruct* which is an inattentive item. *ForcedSex* was more common among adults, but it had a small N (48), so its values may be unstable. *AbsentMinded, Restless, FeltOnTheGo,* and *Wait,* are inattentive and hyperactive items that were easier for adults and were less severe than *DestroydPrprty* and *SetFires.* If we disregard *ForcedSex* because of its instability, there were eight items that differed between youth (4 easier) and adults (4 easier). The reliability estimates were similar for youth, i.e., person r=.87, alpha=.93, and adults, i.e., person r=.86, alpha=.95.

Objective 4. **Test Criterion Validity**

Figure 3 presents the observed relationships of various other scales of the GAIN with the BCS measure. The correlations of the Substance Problem Scale (r=.44), HIV Risk Scale (r=.39), Emotional Problems Scale (r=.59), and Recovery Environment Risk (r=.39) all corresponded with the expectations and were thus supportive of BCS criterion validity. In Figure 3, the BCS gradations formed a corresponding monotonically ascending hierarchy that was supportive of frequency of endorsement as a valid indicator of severity.

Discussion

The *BCS* performed well as a unidimensional measure of the construct of externalizing disorders, with high person reliability and internal consistency estimates. When ADHD items and CD items were treated as separate measures, they were highly correlated; and, after correction for attenuation due to measurement error, the correlation was further improved. This supported the idea that ADHD and CD both assess the same latent construct, i.e., unidimensional measure of externalizing disorders.

Hierarchy

The use of the Rasch measurement model resulted in a very clear hierarchy of externalizing disorders that ascended in severity for inattentive, hyperactive and conduct disorder symptoms. The results indicated that inattentiveness was most common, hyperactivity less common, and severe conduct disorders least common. The implication is also that those with severe CD also tend to have ADHD. In a post hoc analysis of the current data, excluding persons that endorsed 2 or 3 of the low severity CD items (discussed further below), 75% of those with a DSM-IV CD diagnosis also had a DSM-IV diagnosis of ADHD. The overlap of CD and ADHD symptoms is consistent with clinical samples of children which report substantial co-morbidity between ADHD and CD (Biederman et al., 1991; Chan et al., 2008), but the point observed here is that those having higher severity CD symptoms also tended to have ADHD, but not vice versa to the same degree.

Notably, three CD items were in the low severity range. Using DSM-IV criteria, this means that a person that endorses *SkipSchool/Work, LiedConned,* and *StayOut2Late,* gets a diagnosis of CD as does a person that endorses *SetFires, TakeMoneyForce,* and *Weapon.* As noted earlier, this is a problem with categorical classification systems that weight each symptom equally – particularly if the severity specifiers are not used. The implication of the BCS data is that it will be very common for individuals to endorse two or three of the most frequently endorsed CD symptoms, but none or one of the higher CD items. If they endorsed three low severity CD or two low severity CD items and one high severity item, should these patterns qualify as CD? If CD is regarded as more severe than ADHD, then there is a danger that such individuals would be inappropriately regarded as more severely disturbed than they actually are. While the current specifiers might account for this symptom hierarchy, they require subtle

judgments and are rarely used. In a second post hoc analysis, we deleted the 3 low severity CD items to see how much this changed the percentage of persons with CD. Removing the 3 low severity CD items caused the percentage with CD to drop from 51% to 31%. The percentage with ADHD was 43%. Therefore, CD went from being substantially more prevalent than ADHD to substantially less prevalent. We note that since this sample was being screened for SUD, this may also raise the issue of what CD means in populations of substance users, i.e., with the three low severity CD items being very common.

These results indicated that a reinterpretation of CD may be in order for the three most frequently occurring CD items especially in persons being screened for SUD. This problem would be avoided if classification were done using the severity hierarchy. For example, persons scoring between -2 and -1 on the Wright map (Figure 1) could be classified as inattentive even though they might have some low severity CD symptoms as well. Specifically, they could be diagnosed with a low level of externalizing disorders—with the qualifier that they must fit the Rasch model (discussed below).

Also, our findings suggest that if persons have hyperactive symptoms, they will probably have inattentive symptoms if they fit the model. However, having inattentive symptoms alone does not imply the probability that someone will have hyperactivity nor that they will probably have CD. In probabilistic terms, in order to score in the hyperactive range, subjects most likely are endorsing most of the inattentive as well as some hyperactive symptoms.

DIF

The age DIF results showed that youth and adults found equal numbers of items easier to endorse so that these would balance out in the total measure, i.e., unlikely to be biased. *AbsentMinded, Restless, FeltOnTheGo, and Wait,* are ADHD items that were easier for adults

and were less severe than *DestroydPrprty* and *SetFires*, i.e., CD items that were easier for youth. Youth also found *SkipSchool/Work*, a CD item, and *NotFollowInstruct*, an inattentive item, to be easier. Since youth found it somewhat easier to endorse CD items, BCS scores for youth could indicate more CD. This was considered a fine-tuning issue that could be examined in future research.

This may also suggest that some more serious and more age-appropriate conduct disorder items could be created for adults. Another explanation may be that adults might under-report more serious acts of violence or aggression towards others because of more serious consequences. The *SkipSchool/Work* and *NotFollowInstruct* may be inappropriate for adults or may actually be more serious problems when they do occur. While these findings are suggestive of potential item improvement especially for adults, further examination, e.g., Conrad, Dennis, Bezruczko, Funk, & Riley, 2007, of the effect on scores was beyond the scope of this study.

Criterion Validity

The criterion validity correlation results indicated a pattern of association with variables that are frequently co-morbid with externalizing disorders, e.g., substance use disorders, emotional problems, and other adverse life events that would be expected with increased severity of externalizing disorders. This supports the interpretation of the hierarchy as representing severity of the disorder in terms of increasing life problems in general. What may be most significant about this finding is the strong correspondence of the severity groupings with the categories of inattentive, hyperactive and conduct disorders. This suggests that classification could correspond well with cut-points along the measurement spectrum, but further validation of this observation is clearly needed.

Item Improvement Issues

The item *ForcedSex* appeared unstable due to a low number of endorsements. Evaluation of alternative wording, e.g., using words such as "seduced or pressured against their will," may be useful. Another item that could be updated is *StolStorBadChecks* which may be revised to include other financial behavior such as ATM fraud and credit card theft. Indeed, qualitative work would be useful to clarify any wording issues and expand the pool of items, especially for adults. The item findings were viewed as "fine tuning" issues that could improve the BCS but did not threaten its general usefulness at this time.

Limitations

It should be noted that the GAIN's BCS only focuses on ADHD and CD. Other researchers already include or may want to include other externalizing disorders, including Axis I disorders with childhood onset (e.g., oppositional defiant disorder, intermittent explosive disorder), Axis I disorders with adult onset (e.g., pathological gambling, other impulse control disorders), Axis II personality disorder from cluster B (e.g., antisocial personality disorder, borderline personality disorder), and/or even substance disorders (e.g., abuse, dependence). We can however point out that the BCS approach has demonstrated excellent structural fit to the data (CFI=.87, RMSEA=0.05) in confirmatory factor analysis (Dennis, Chan & Funk, 2006) and in selected samples has been shown to correlate with measures of pathological gambling and personality disorders (Rush, Dennis, Scott, Castel, & Funk, 2008).

The BCS is a self-report measure and has all the limitations of self-report such as possibility of socially desirable responses. Cross-validation with clinician assessments and other collateral informants would further test the validity of the BCS. The sample of persons being screened for substance abuse was large and appropriate, but its generalizability should be tested in samples that might be less severe, such as students, or more severe, such as prisoners. We also

note that all measures used for criterion validation were, like the BCS, part of the GAIN so that shared method variance may have tended to increase correlations. Further criterion validation using external measures using a variety of methods and settings should be conducted. The sample, while being large and diverse in many ways, was from persons being screened for substance use disorders. Therefore, the study should be replicated on more diverse samples as well.

Finally, while one of the largest samples available to date, the GAIN data set is not a random sample of all people entering the substance abuse treatment and thus may not be representative of them. Clinical samples are themselves characterized by much higher rates of co-morbidity than one would normally see in community settings for the same disorders (Kessler, Chiu, Demler, Merikangas, & Walters, 2005).

Conclusion

In general, the BCS functioned well in this sample of persons being assessed for substance abuse and was found to be valid across age groups. However, several items showed potential for improvement with future qualitative work. Further research on the severity hierarchy, including the impact of differential item severity by age, will be required to validate these findings beyond the current study population.

The Rasch severity hierarchy of attention deficit to hyperactivity to conduct disorders provided a perspective that suggested that a dimensional measure could be used as an alternative and, in some ways, as an improvement, to categorical diagnosis and common dimensional approaches. The finding of the three low severity conduct disorder items may lead to revising categorical criteria, especially in substance use disorders, and developing a dimensional severity hierarchy. It illustrates how the Rasch measurement model may be used to suggest improvements

in the assessment of externalizing disorders and demonstrates its potential for examining other psychological disorders as well.

References

Achenbach, T.M. (1991). *Manual for the Teacher's Report Form and 1991 Profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., & Edelbrock, C. S. (1978). The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin, 85*, 1275– 1301.

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

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision.* Washington, DC, American Psychiatric Association.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, 22(1), 1145-1146. http://www.rasch.org/rmt/rmt221.pdf

Barkley, R., Fischer, M., Smallish, L., & Fletcher, K. (2004). Young adult follow-up of hyperactive children: Antisocial activities and drug use. *Journal of Child Psychology and Psychiatry*, *45*(2), 195-211.

Biederman, J., Newcorn, J., & Sprich S. (1991). Comorbidity of attention deficit hyperactivity disorder with conduct, depressive, anxiety, and other disorders. *American Journal of Psychiatry*, *148*(5), 564-577.

Biederman, J., Wilens, T., Mick, E., Faraone, S., & Spencer, T. (1998). Does attention-deficit hyperactivity disorder impact the developmental course of drug and alcohol abuse and dependence? *Biological Psychiatry, 44*(4), 269-273.

Bond, T.G. & Fox, C.M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences.* (2nd Ed.) Mahwah, NJ: Erlbaum Associates.

Boyle, M. H., Offord, D. R., Racine, Y., Szatmari, P., Fleming, J. E. & Sanford, M. (1996). Identifying thresholds for classifying childhood psychiatric disorder: Issues and prospects. *Journal of the American Academy of Child and Adolescent Psychiatry, 35,*1440–1448.

Buitelaar, J. (2007, February). *Diagnosis of ADHD in adults.* Paper presented at American Psychiatric Association DSM-V Planning Conference on Externalizing Disorders of Childhood, Mexico City. http://www.dsm5.org/Research/Pages/ExternalizingDisordersofChildhood%28Attention-deficitHyperactivityDisorder,ConductDisorder,Oppositional-DefiantDisorder,Juven.aspx

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56: 81-105.

Chan, Y., Dennis, M., & Funk, R. (2008). Prevalence and comorbidity co-occurrence of major internalizing and externalizing disorders among adolescents and adults presenting to substance abuse treatment. *Journal of Substance Abuse Treatment, 34*(14), 14-24.

Chestnut Health Systems. GAIN: Administrative guide for the GAIN and related measures. Author. [Version 5.6.2]. 2010. Bloomington, IL.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates**.**

Cohen, N. J., Gotlieb, H., Kershner, J., & Wehrspann, W. *(*1985*).* Concurrent validity of the internalizing and externalizing profile patterns for the Achenbach Child Behavior Checklist. *Journal of Consulting and Clinical Psychology, 53,* 724–728.

Conners, C. K., Erhardt, D., & Sparrow, E. (1999). *Conners' Adult ADHD Rating Scales (CAARS)*. North Tonawanda, NY: Multi-health Systems.

Cohen, N., Gotlieb, H., Kershner, J., & Wehrspann, W. (1985). Concurrent validity of the internalizing and externalizing profile patterns of the Achenbach Child Behavior Checklist. *Journal of Consulting and Clinical Psychology, 53*(5), 724-728.

Conrad, K. J., Conrad, K. M., Dennis, M. L., Riley, B.B., & Funk (2009). Validation of the HIV Scale to the Rasch Measurement Model, GAIN Methods Report 1.1. Chicago, IL: Chestnut Health Systems. http://www.chestnut.org/li/gain/psychometric_reports/Conrad_et_al_2009_HIV_Rasch_Report.pdf

Conrad, K.J., Dennis, M.L., Bezruczko, N., Funk, R., & Riley, B. (2007). Substance use disorder symptoms: Evidence of differential item functioning by age. *Journal of Applied Measurement, 8*(4), 373-387.

Conrad, K.J. & Smith, E.V. (2004). Applications of Rasch analysis in health care. *Medical Care* 42 (Suppl I).

Dennis, M. L., Dawud-Noursi, S., Muck, R. D., & McDermeit, M. (2003). The need for developing and evaluating adolescent treatment models. In S. J. Stevens, & A. R. Morral (Eds.). *Adolescent substance abuse treatment in the United States: Exemplary models from a national evaluation study (pp. 3 –34).* Binghamton, NY: Haworth Press.

Dennis, M. L., Chan, Y-F., & Funk, R. R. (2006). Development and validation of the GAIN short screener (GSS) for internalizing, externalizing, and substance use disorders and crime/violence problems among adolescents and adults. *The American Journal on Addictions, 15,* 80-91.

Dennis, M. L. (n. d.). GAIN home page. http://www.chestnut.org/LI/gain/

Disney, E., Elkins, I., McGue, M., Iacono, W. (1999) Effects of ADHD, conduct disorder, and gender on substance use and abuse in adolescence. *American Journal of Psychiatry, 156,* 1515–1521.

Draba, R.E. (1977). The identification and interpretation of item bias (*Research Memorandum No. 26*). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Edwards, A.C. & Kendler, K.S. (2012). Twin study of the relationship between adolescent attention-deficit/hyperactivity disorder and adult alcohol dependence. *Journal of Studies on Alcohol and Drugs, 73,* 185–194.

Elder, C., McNamara, T. & Congdon, P. (2003). Rasch techniques for detecting bias in performance assessment: An example comparing the performance of native and non-native speakers on a test of academic English. *Journal of Applied Measurement,* 4, 181-197.

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

Epstein, J. & Kollins, S. (2006) Psychometric properties of an adult ADHD diagnostic interview. *Journal of Attentive Disorders, 9,* 504-514.

Erhardt, D., Epstein, J. N., Conners, C. K., Parker, J. D. A., & Sitarenios, G. (1999). Self-ratings of ADHD symptoms in adults II: Reliability, validity, and diagnostic sensitivity. *Journal of Attention Disorders, 3,* 153-158.

Frick, P. (2006). Developmental pathways to conduct disorder. *Child and Adolescent Psychiatric Clinics of North America, 15*(2), 311-331.

Guttmannova, K., Szanyi, J.M., Cali, P.W. (2008). Internalizing and externalizing behavior problem scores cross-ethnic and longitudinal measurement invariance of the behavior problem index. *Educational and Psychological Measurement, 68*(4), 676-694.

Hinshaw, S.P. & Nigg, J.T. (1999). Behavior rating scales in the assessment of disruptive behavior problems in childhood. In D. Shaffer, C.P. Lucas, & J.E. Richter (Eds.), *Diagnostic Assessment in Child and Adolescent Psychopathology* (p. 91-126). New York: Guilford Press.

Jensen, P.S., Martin, D., & Cantwell, D.P. (1997). Comorbidity in ADHD: Implications for research, practice, and DSM-V. *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 1065-1079.

Kessler, R.C., Chiu, W.T., Demler, O., Merikangas, K.R., Walters, E.E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62(6), 617-27.

Kim-Cohen, J., Caspi, A., Moffitt, T. E., Harrington, H., Milne, B. J. & Poulton, R. (2003). Prior juvenile diagnoses in adults with mental disorder: Developmental follow-back of a prospective longitudinal cohort. *Archives of General Psychiatry, 60,* 709–717.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry, 56* (10), 921-926.

Krueger, R.F. & Markon, K.E. (2008). Understanding psychopathology: Melding behavior genetics, personality, and quantitative psychology to develop an empirically based model. *Current Directions in Psychological Science,* 15(3):113.

Krueger, R., Markon, K., Patrick, C.J., & Iacono, W. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for DSM-V. *Journal of Abnormal Psychology,* 114(4), 537-550.

Krueger, R. F. & South, S. C. (2009). Externalizing disorders: Cluster 5 of the proposed meta-structure for DSM-V and ICD-11. *Psychological Medicine,* 39, 2061-2070.

Lennox, R.D., Dennis, M.L., Scott, C.K., Funk, R. (2006). Combining psychometric and biometric measures of substance use. *Drug & Alcohol Dependence,* 83, 95–103.

Lilienfeld, S.O. (2003). Comorbidity between and within childhood externalizing and internalizing disorders: Reflections and directions. *Journal of Abnormal Child Psychology, 31(3), 285–291*.

Linacre, J.M. (2011). *Winsteps Rasch Measurement* (Version 3.72.0) [Computer Software]. Available from http://www.winsteps.com.

Linacre, J.M. (1998a). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions,* 1(2), 636.

Linacre, J.M. (1998b). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2(3),* 266-283.

Linacre, J.M. & Wright, B.D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions,* 1994, 8(2), 360. http://www.rasch.org/rmt/rmt82a.htm

Loeber, R., Burke, J. D., Lahey, B. B., Winters, A. & Zera, M. (2000). Oppositional defiant and conduct disorder: A review of the past 10 years, Part I. *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 1468–1484.

Longford, N.T., Holland, P.W., & Thayer, D.T. (1993). Stability of the MH D-DIF statistics across populations. In P.W. Holland & H.Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lunz, M. (2010, January). Using the very useful Wright map. *Measurement Research Associates: Test Insights.* http://www.rasch.org/mra/mra-01-10.htm

Markon, K.E., Chmielewski, M., & Miller, C.J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin.* Advance online publication. Doi: 10.1037/a0023678.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50, 741-749.

Mezzacappa, E. (2007, February). *Item response theory applied to parent ratings of aggression and delinquency.* Paper presented at American Psychiatric Association DSM-V Planning Conference on Externalizing Disorders of Childhood, Mexico City. http://www.dsm5.org/Research/Pages/ExternalizingDisordersofChildhood%28Attention-deficitHyperactivityDisorder,ConductDisorder,Oppositional-DefiantDisorder,Juven.aspx

Modisette, K. C., Hunter, B. D., Ives, M. L., Funk, R. R. & Dennis, M. L. (2009). NORMS including alpha, mean, N, sd, ICC for Adolescents (by demographics) and overall for Young Adults (18-25) and Adults (18+) using the CSAT 2008 V5 Dataset [Electronic version]. Normal, IL: Chestnut Health Systems. http://chestnut.org/LI/Posters/Norms_by_Demog_2008dataset.xls

Nock, M., Kazdin, A., Hiripi, E., & Kessler, R. (2006). Prevalence, subtypes, and correlates of DSM-IV conduct disorder in the National Comorbidity Survey Replication. *Psychological Medicine, 26,* 699-710.

Norman, G.R., Sloan, J.A., & Wyrwich, K.W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care, 41(5),* 582-592.

Paek, I. (2002). *Investigations of differential Item functioning: Comparisons among approaches, and extension to a multidimensional context.* Unpublished doctoral dissertation, University of California, Berkeley.

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relationships, and behavior problems in children. *Journal of Marriage and Family*, 48, 295-307.

Rapport, M., LaFond, S., Sivo, S., (2009). Unidimensionality and developmental trajectory of aggressive behavior in clinically-referred boys: A Rasch analysis. *Journal of Psychopathology and Behavioral Assessment.*

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedogogiske Institut. (Republished Chicago: The University of Chicago Press: 1980).

Rush, B.R., Dennis, M.L., Scott, C.K., Castel, S., and Funk, R. (2008). The interaction of co-occurring mental disorders and recovery management checkups on substance abuse treatment participation and recovery. *Evaluation Review,* 32; 7-38.

Sher, K.J. & Trull, T.J. (1994). Personality and disinhibitory psychopathology: Alcoholism and antisocial personality disorder. *Journal of Abnormal Psychology*, 103, 92–102

Scheuneman, J.D. and Subhiyah, R.G. (1998). Evidence for the validity of a Rasch technique for identifying differential item functioning. *Journal of Outcome Measurement*, 2, 33-42.

Simonoff, E., Elander, J., Holmshaw, J., Pickles, A., Murray, R. & Rutter, M. (2004). Predictors of antisocial personality: Continuities from childhood to adult life. *British Journal of Psychiatry, 184*, 118–127.

Smith, E.V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3(2),* 205-231.

Stein, M.A. & O'Donnell, J.P. (1985). Classification of children's behavior problems: Clinical and quantitative approaches. *Journal of Abnormal Child Psychology*, 13, 269-280.

Szatmari, P., Boyle, M., and Offord, D.R. (1989). ADDH and conduct disorder: Degree of diagnostic overlap and differences among correlates. *Journal of American Academy of Child and Adolescent Psychiatry, 28*(6), 865-872.

Vollebergh, W. A., Iedema, J., Bijl, R.V., deGraaf, R., Smit, F., Ormel, J. The structure and stability of common mental disorders: the NEMESIS study. *Archives of General Psychiatry*, 58(6), 597-603.

Wang, W.C. (2000). Modeling effects of differential item functioning in polytomous items. *Journal of Applied Measurement, 1*, 63-82.

Wilens, T., Biederman, J., & Spencer, T. (2002). Attention deficit-hyperactivity disorder across the lifespan. *Annual Review of Medicine, 53,* 113-131.

Wilens, T. (2004). Attention-deficit/hyperactivity disorder and the substance use disorders: The nature of the relationship, subtypes at risk, and treatment issues. *Psychiatric Clinics of North America, 27,* 283-301.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum Associates.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-059). Beverly Hills, CA: Sage.

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago: University of Chicago, MESA Press.

Table 1. Behavioral Complexity Scale Items, Labels, Measures and Fit Statistics

BCS Item Descriptors	**Item Label**	**Measure**	**Infit**	**Outfit**
Inattentive Disorder Scale				
1. Made mistakes because you were not paying attention.	MistakesAttn	-2.60	.99	1.00
2. Had a hard time paying attention at school, work or home.	HardPayAttnSchool	-1.47	.90	.83
3. Had a hard time listening to instructions at school, work or home.	ListenInstructins	-.99	.88	.83
4. Not followed instructions or not finished your assignments.	NotFollowInstruct	-1.65	.95	.91
5. Had a hard time staying organized or getting everything done.	StayOrg	-1.14	.97	.99
6. Avoided things that took too much effort, like school work or paperwork.	AvoidEffort	-1.12	.92	.87
7. Lost things that you needed for school, work or home.	LostThings	-.91	.99	.97
8. Been unable to pay attention when other things were going on.	UnablePayAttnThingsGo	-1.29	.87	.81
9. Been forgetful or absentminded.	Absentminded	-1.47	.93	.90
Hyperactivity Impulsivity Scale				
10. Fidgeted or had a hard time keeping your hands or feet still when you were supposed to.	Fidget	-.53	.94	.90
11. Been unable to stay in a seat or where you were supposed to stay.	StaySeated	.18	.89	.83
12. Felt restless or the need to run around or climb on things.	Restless	.51	.94	.91
13. Gotten in trouble for being too loud when you were playing or relaxing.	LoudPlay	-.22	.97	.94
14. Felt like you were always on the go or driven by a motor.	FeltOnTheGo	.20	1.02	1.04
15. Talked too much or had others complain that you talked too much.	TalkTooMuch	.47	1.08	1.21
16. Gave answers before the other person finished asking the question.	AnswrB4Question	-1.06	1.11	1.17
17. Had a hard time waiting for your turn.	Wait	-.30	.93	.90
18. Interrupted or butted into other people's conversations or games.	Interrupted	-.44	1.04	1.06
Conduct Disorder Scale				

19. Been a bully or threatened other people.	Bully	.13	1.01	1.02
20. Started fights with other people.	StartFights	.36	1.05	1.08
21. Used a weapon in fights.	Weapon	1.75	1.05	1.33
22. Been physically cruel to other people.	PhysCrulPeopl	.87	.98	1.01
23. Been physically cruel to animals.	PhysCrulAnmal	3.00	1.09	1.31
24. Taken a purse, money or other things from another person by force.	TakeMoneyForce	2.30	.99	.94
25. Forced someone to have sex with you when they did not want to.	ForcedSex	5.11	1.03	1.16
26. Set fires.	SetFires	2.04	1.07	1.05
27. Broken windows or destroyed property.	DstroydPrprty	.79	.99	1.02
28. Taken money or things from a house, building or car.	TakeMoneyHome	.85	1.00	1.02
29. Lied or conned to get things you wanted or to avoid having to do something.	LiedConned	-1.19	.99	.94
30. Taken things from a store or written bad checks to buy things.	StolStorBadChks	.62	1.04	1.17
31. Stayed out at night later than your parents or partner wanted.	StayOut2Late	-2.02	1.12	1.21
32. Run away from home (partner) for at least one night.	RunAwayOvrnite	.40	1.11	1.21
33. Skipped work or school.	SkipSchool	-1.17	1.21	1.38

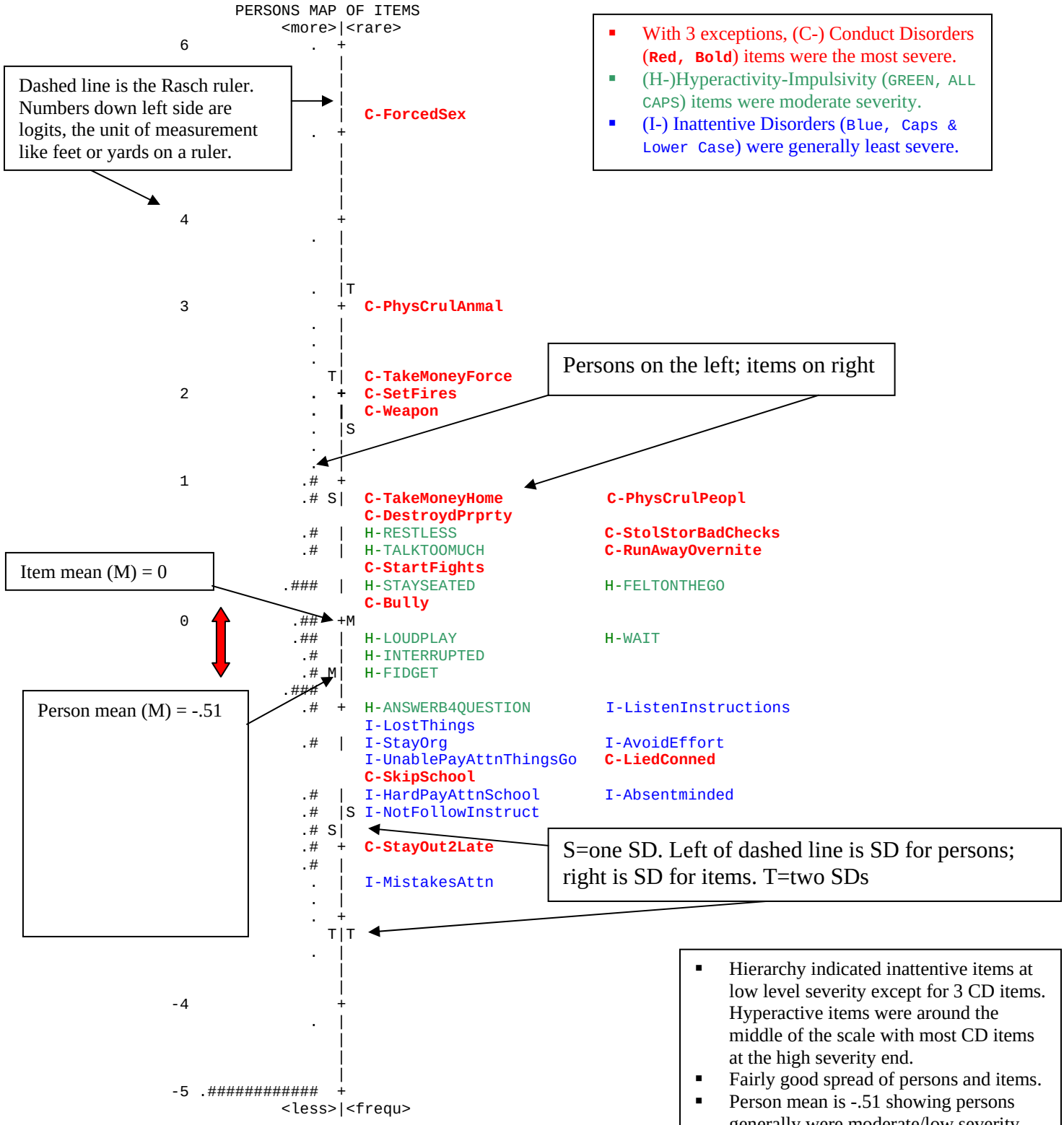

Figure 1. Wright Map.

Figure 2. BCS Differential Item Functioning (DIF) by age.

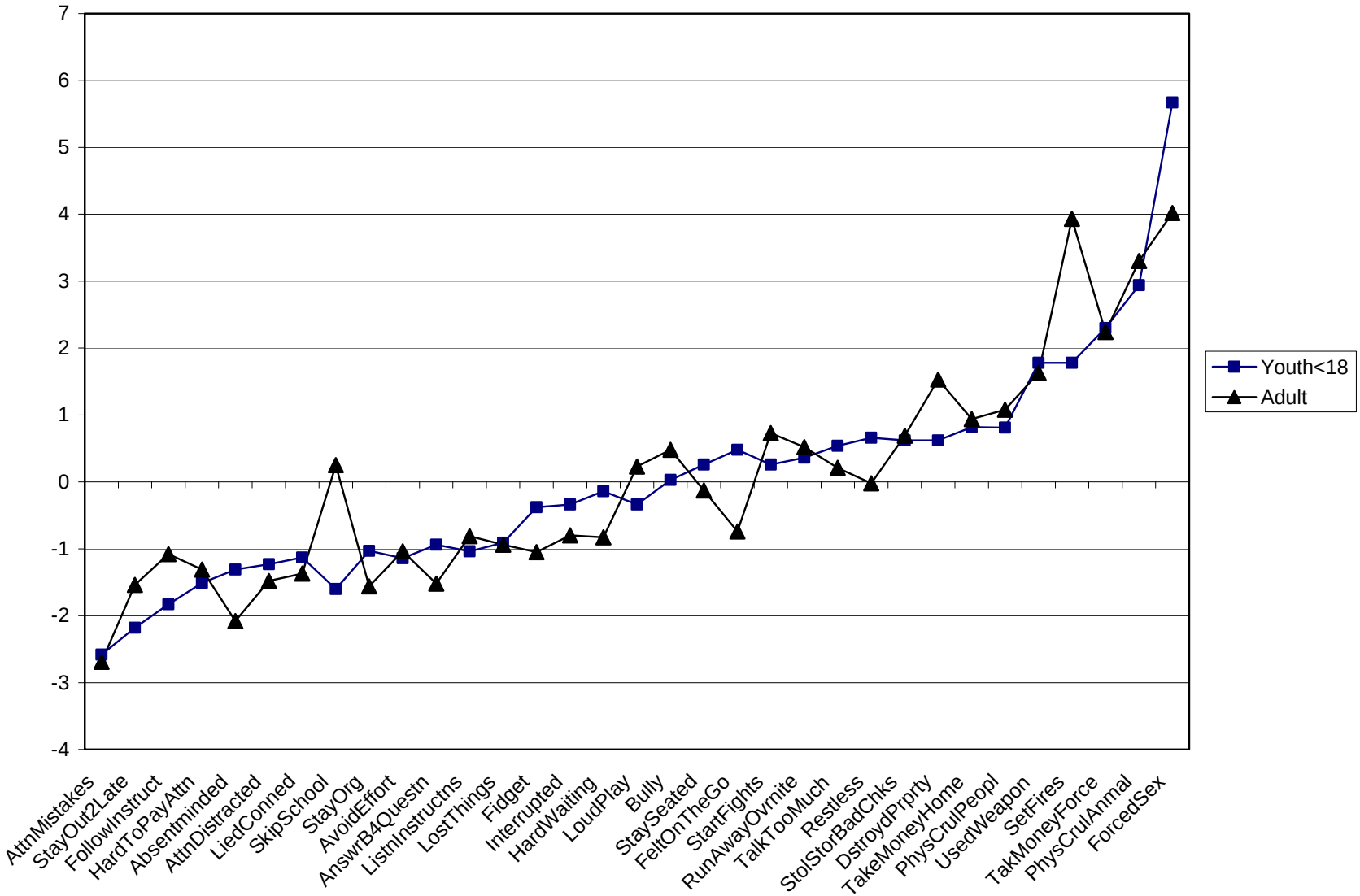

Figure 3. Correlations of Rasch BCS and performance of hierarchy against four criterion variables.