Published in final edited form as:

Psychol Assess. 2011 March ; 23(1): 245–261. doi:10.1037/a0021501.

Bifactor and Item Response Theory Analyses of Interviewer Report Scales of Cognitive Impairment in Schizophrenia

Steven P. Reise¹, Joseph Ventura, Ph.D.², Richard S. E. Keefe, Ph.D.³, Lyle E. Baade, Ph.D.⁴, James M. Gold, Ph.D.⁵, Michael F. Green, Ph.D.^{2,6}, Robert S. Kern, Ph.D.^{2,6}, Raquelle Mesholam-Gately, Ph.D.⁷, Keith H. Nuechterlein, Ph.D.^{2,1}, Larry J. Seidman, Ph.D.⁷, and Robert Bilder, Ph.D.^{2,1}

¹UCLA Department of Psychology ²Semel Institute for Neuroscience & Human Behavior, Geffen School of Medicine at UCLA ³Department of Psychiatry and Behavioral Sciences, Duke University, Durham, North Carolina ⁴University of Kansas School of Medicine at Wichita, Kansas ⁵Maryland Psychiatric Research Center, University of Maryland School of Medicine, Baltimore, Maryland ⁶VA Greater Los Angeles Healthcare System, Los Angeles, California ⁷Massachusetts Mental Health Center, Public Psychiatry Division of the Beth Israel Deaconess Medical Center Department of Psychiatry, Harvard Medical School, Boston, Massachusetts

Abstract

We conducted psychometric analyses of two interview-based measures of cognitive deficits: the 21-item Clinical Global Impression of Cognition in Schizophrenia (CGI-CogS; Ventura et al., 2008), and the 20-item Schizophrenia Cognition Rating Scale (SCoRS; Keefe et al., 2006), which were administered on two occasions to a sample of people with schizophrenia. Traditional psychometrics, bifactor analysis, and item response theory (IRT) methods were used to explore item functioning, dimensionality, and to compare instruments. Despite containing similar item content, responses to the CGI-CogS demonstrated superior psychometric properties (e.g., higher item-intercorrelations, better spread of ratings across response categories), relative to the SCoRS. We argue that these differences arise mainly from the differential use of prompts and how the items are phrased and scored. Bifactor analysis demonstrated that although both measures capture a broad range of cognitive functioning (e.g., working memory, social cognition), the common variance on each is overwhelmingly explained by a single general factor. IRT analyses of the combined pool of 41 items showed that measurement precision is peaked in the mild to moderate range of cognitive impairment. Finally, simulated adaptive testing revealed that only about 10 to 12 items are necessary to achieve latent trait level estimates with reasonably small standard errors for most individuals. This suggests that these interview-based measures of cognitive deficits could be shortened without loss of measurement precision.

Keywords

item response theory; CGI-CogS; SCoRS; schizophrenia and cognitive deficits; computerized adaptive testing

Correspondence concerning this article should be addressed to Steven P. Reise, Department of Psychology, UCLA, Los Angeles, CA 90095, reise@psych.ucla.edu.

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/PAS

Cognitive deficits associated with schizophrenia are seen by academia and industry as a potential indication for new drug treatment development (Green, 2007; Green and Nuechterlein, 2004; Hyman and Fenton, 2003). New instrumentation is needed to evaluate possible efficacy of new treatments, particularly to obtain “co-primary” endpoint data that can complement objective cognitive test results. Moreover, patient reported outcomes (PROs) are increasingly important in clinical trials research, and the Food and Drug Administration (FDA) has identified a development cycle for such instruments that encourages the adoption of strong psychometric methods (U.S. Department of Health and Human Services, FDA, February, 2006). In this study, we conduct psychometric analyses of two semi-structured interview-based measures of cognitive deficits in schizophrenia. Specifically, we analyze the 21-item CGI-CogS (Ventura et al., 2008), and the 20-item SCoRS (Keefe et al., 2006). These measures were selected for analyses and possible combining because: a) they were the only two interview-based measures specifically designed as co-primary measures to assess cognitive deficits in schizophrenia, b) contain similar item content, and c) each ostensibly attempts to measure the same construct.

The overarching goals of the present analyses are as follows. Our first objective is to compare the two instruments on a number of criteria, such as average inter-item correlation, dimensionality, coherence of factor structure, degree of saturation by a general factor, and floor or ceiling effects at the item level. As described below, the item content of both measures covers a broad range of cognitive domains (e.g., working memory, visual learning, problem solving) thought to be important in properly assessing cognition. We will use both bifactor (see, Schmid & Leiman, 1957; Chen, West, & Sousa, 2006; Reise, Morizot, & Hays, 2007; Immekus & Imbrie, 2008) and more commonly used exploratory factor analytic techniques (e.g., principal axis extraction, oblimin rotation) to investigate the extent to which item responses on each instrument are saturated by a general factor, and to what degree meaningful multidimensionality can be identified. The dimensionality issue is critical, given recent initiatives to assess and track change on separate dimensions of cognitive functioning.

If the item responses to each measure are dominated by a single general factor, we will then fit a unidimensional item response theory (IRT; Embretson & Reise, 2000) model to the responses from the combined measures. The objectives of fitting the IRT model are twofold. First, we want to systematically explore how category ratings are related to a common latent variable and where on the latent trait continuum each item provides the most discrimination among individuals. A second objective is to use IRT modeling to explore how many items are needed for precise measurement of cognitive deficits. This latter issue is also critical given the enormous time and expense of clinical interviewer-based measures and the need for short measures that yield reasonably precise scores in clinical trial, intervention, and developmental research.

Methods

The MATRICS Project

The data for the present study are drawn from the MATRICS (measurement and treatment research to improve cognition in schizophrenia) Psychometric and Standardization Study (PASS) led by Co-Chairs of the Neurocognition Committee (Drs. Nuechterlein and Green). The multi-site MATRICS project was funded by the National Institute of Mental Health (NIMH) and involved extensive interactions among academic, NIMH, FDA, and pharmaceutical industry representatives (Nuechterlein et al., 2008; Green and Nuechterlein, 2004; Green et al., 2008). The overarching goal of MATRICS – PASS was to develop and evaluate a battery of cognitive tests that could serve to stimulate the development and

evaluation of new drugs to improve cognition in schizophrenia. To accomplish this objective, MATRICS PASS researchers sought to identify a set of measures with good psychometric properties and would be repeatable in clinical trial contexts (i.e., a set of "consensus cognitive performance measures").

One of the chief goals underlying the MATRICS – PASS project was the FDA recommendation that for any new pharmaceutical agents seeking approval as cognitive enhancers, improvements on objective cognitive batteries would be necessary, but not sufficient. According to FDA officials, additional evidence also is needed to corroborate that objectively measured change is, in fact, clinically meaningful and relevant to "real world" social, educational, and occupational functioning (Green et al., 2008). As a result, additional "co-primary" measures (e.g., SCoRS, CGI-CogS) would need to be included along with objective test results in clinical trials of cognition enhancing medications that were seeking FDA approval.

Participants

The sample of participants for the MATRICS – PASS project was selected to be representative of stable patients with schizophrenia or schizoaffective disorder and who would be typical of patients who participate in clinical trial research. The participants can be described as a sample of chronic outpatients with reasonable levels of adult daily living skills. Although the sample is limited in terms of individuals with severe levels of cognitive impairment (see results below), many individuals are included who are quite limited in terms of educational and occupational functioning.

Specifically, participants were selected according to the following criteria: diagnosis of schizophrenia or schizoaffective disorder, depressed type, based on diagnostic interview; no medication changes in previous month and none expected for the following month; stable outpatient or rehabilitation center status; age 18–65 years; no substance dependence in past 6 months; no substance abuse in past month; no clinically significant neurological disease or head injury as determined by medical history; ability to understand spoken English sufficiently to complete testing procedures; and ability to comprehend the consent form appropriately. We also excluded individuals who did not meet the substance use or dependence exclusion criteria but who had a clearly excessive amount of lifetime alcohol or drug consumption over a 10-year period or had been using alcohol or drugs heavily in the 3 days prior to testing (Nuechterlein et al., 2008).

In MATRICS – PASS, data were collected on two occasions four weeks apart between January 2004 and August 2005. At baseline, the number of participants was 176. At one-month follow-up, nine individuals were lost. Combining baseline and follow-up data, the total number of possible data points for analyses was 343.1 However, not all individuals were able to complete both the CGI-CogS and SCoRS at follow-up. Analyses in this paper thus are based on sample sizes that vary somewhat across analyses. Specifically, the number of data points for within scale descriptive statistics and factor analyses are 319 for CGI-Cogs and 323 for SCoRS. The subsequent item response theory and simulated computerized adaptive testing analyses are based on 315 data points that had no missing data on either instrument. The raw aggregated scores derived from the CGI-CogS and SCoRS had 4-week test-retest reliability estimates of 0.80 and 0.82, respectively.

¹Our data consists of the same people measured twice and thus the present item response data are obviously not independent across time. Although problems with conducting null hypotheses tests in the presence of violation of independence are well recognized (i.e., standard errors are too low), violating independence has zero biasing effect on the parameter estimates. Thus, factor loadings and item response theory model discrimination parameters are valid estimates.

At baseline, the MATRICS participants ($n = 176$) had a mean age of 44.0 (11.2) years, a mean education of 12.4 (2.4), and were 76% male. The marital status was as follows; single 61.4%, married or co-habiting 9.7%, divorced 23.9%, widowed 2.3% and, separated 2.8%. The ethnicity and racial distribution was as follows; Caucasian (59%), African American (29%), Hispanic/Latino (6%), Asian Pacific (<1%), and Native American/Alaskan (4%). Diagnosis according to DSM-IV was Schizophrenia (86%) or Schizoaffective, depressed type (14%), with Brief Psychiatric Rating Scale (BPRS; Ventura, Green, Shaner, & Liberman, 1993) positive symptoms minimum = 3; maximum = 19; mean = 7.7 (3.77) and negative symptoms minimum=3; maximum=14; mean = 5.98 (2.56). These BPRS scores are interpreted as indicating that participants are in the mild range of symptomatology and on average below clinically significant levels. This is consistent with guidelines for the optimal design of clinical trials to enhance cognitive function in schizophrenia, and specifically with the recommendation that these should include individuals who are clinically stable and have no more than moderate delusions, hallucinations, or formal thought disorder symptoms (Buchanan et al., 2005).

Assessments

The collection of data on both the SCoRS and CGI-CogS instruments is complicated and thus we briefly describe how the item responses analyzed herein were obtained. When a participant arrives for an assessment, a previously trained clinician conducts an in-person semi-structured interview centered on the CGI-CogS and SCoRS items (see description below). The clinician then rates each item based on this interview. In addition, all participants are asked to provide the name of an "informant" (e.g., family member, friend, social worker) that knows them well. The research team then contacts the informant and the clinician conducts the same semi-structured interview either in person or over the phone, and ultimately rates each CGI-CogS and SCoRS item based on this second source of information. At no time do participants or informants directly rate themselves on each item. Finally, based on these two sources of information combined with his or her judgment, the clinician makes a final rating on each item. It is these rater item ratings that comprise the data for this study. The two measures of cognitive deficits were counterbalanced between, but not within individuals; that is, each individual received the same order at baseline and retest, but order differed across individuals.

Clinical Global Impression of Cognition in Schizophrenia (CGI-CogS)

The 21-item CGI-CogS (Ventura et al., 2008) has a general background section and includes four major categories for evaluation: 1) activities of daily living, 2) severity of impairment for cognitive domain, 3) global severity of cognitive impairment, and 4) global assessment of functioning. The CGI-CogS items were written to parallel the cognitive impairments assessed by established batteries of objective cognitive tests (see Nuechterlein et al., 2004). Specifically, the item content reflects cognitive deficits selected from the seven domains of functioning included in the MATRICS Consensus Cognitive Battery. These domains are: 1) working memory (2 items), 2) attention/vigilance (3 items), 3) verbal learning and memory (2 items), 4) spatial learning (3 items), 5) reasoning and problem solving (4 items), 6) speed of processing (3 items), and 7) social cognition (4 items). In the interview, the items are ordered by these content domains.

Each item on the CGI-CogS is phrased as series of questions (i.e., prompts) that attempt to provide concrete examples for each concept assessed. For example, the concept underlying Item #1 is "difficulty maintaining newly learned verbal information in mind for brief periods (long enough to use)." The actual question (prompt) is: "Do you forget names of people you just met? Do you have trouble recalling telephone numbers you hear? Do you have trouble remembering what your Dr. just said during visits? Do you find you need to write down

information to remember?" As the participant or informant responds and describes their functioning in these areas in the last month, the rater both makes a numeric rating, and writes down specific concrete examples. The CGI-CogS uses a 7-point Likert scale for its ratings: 1) not at all impaired, 2) minimal, cognitive deficits but functioning is generally effective, 3) mild, cognitive deficits with some consistent effect on functioning, 4) moderate, cognitive deficits with clear effects on functioning, 5) serious, cognitive deficits which interfere with day-to-day functioning, 6) severe, cognitive deficits that jeopardize independent living, and 7) cognitive deficits are so severe as to present a danger to self and others. Items can also be scored N/A for not applicable or insufficient information. The cognitive deficit assessed by each is displayed in Table 1 and the content domain for each item is shown in Table 3.

Schizophrenia Cognition Rating Scale (SCoRS)—The SCoRS (Keefe et al., 2006) is a 20-item interview-based measure of cognitive deficits and the degree to which they affect day-to-day functioning. The items were developed by a team of content experts to assess a variety of cognitive domains (i.e., memory, attention, problem solving, working memory, learning, speed of processing, and social cognition) that were chosen because of their severity of impairment in many patients with schizophrenia and the empirically established relationship of these cognitive deficits to impairments in aspects of functional outcome. Although these deficits were not originally developed to exactly match the seven cognitive domains included in the MATRICS Consensus Cognitive Battery, the item content on SCoRS is very similar to CGI-CogS (see Keefe et al. 2006, Ventura et al. 2008, as well as Ventura et al., under review). The original purpose of the SCoRS was to "serve as a co-primary measure in clinical trials of cognitive enhancing drugs for schizophrenia" (Keefe, et al., 2006, p. 426).

The format for each SCoRS item differs from CGI-CogS in several ways. First, SCoRS items begin with "do you have difficulty...." and then target a more specific cognitive/behavioral phenomenon. For example, SCoRS Item #1 is "Remembering names of people you know or meet? Example: roommate, nurse, doctor, family & friends" and Item #2 is "remembering how to get places? Example: friend's house, store, restroom, own room." Second, in contrast to CGI-CogS's 7-point scale, SCoRS items are rated on a 4-point scale: none, mild, moderate, and severe impairment. However, each item has a unique anchor for each of the 1 to 4 ratings. For example, on Item #1, anchors are: 1) no impairment, 2) mild = sometimes forgets names, 3) moderate = frequently forgets names, and 4) severe = almost always forgets names. For Item #2, anchors are: 1) no impairment, 2) mild = sometimes forgets how to get places, 3) moderate = only able to get to very familiar places, and 4) severe = unable to get any place without assistance because of difficulty with memory. In other words, the anchor points for each SCoRS item describe the frequency ("sometimes", "almost always") to which a deficit impairs a specific aspect of day-to-day functioning. The underlying concept ostensibly assessed by the SCoRS items is displayed in Table 2 and the content domain is listed in Table 4.

Traditional Item Statistics

Tables 1 and 2 display traditional item psychometrics for both measures separately. Indices include the item mean, standard deviation, corrected item-total correlations, and the response proportions in each category. These indices point to major differences between the two instruments. Most notably, the item-test correlations are strikingly higher on CGI-CogS relative to SCoRS. Moreover, coefficient alpha for CGI-CogS item responses is .95 with an average inter-item correlation of .47 (range from .15 to .65). The coefficient alpha for SCoRS item responses is .89 with an average inter-item correlation of .29 (range from .09 to .53). The difference in the average inter-item correlation is striking and could be taken to

suggest that the two instruments are measuring cognitive deficits at different levels of generalization, with the CGI-CogS having more homogeneous item content relative to SCoRS. However, the raw summed scores on the two scales are correlated $r = .84$, a value which, when corrected for attenuation due to unreliability in observed scores is $r = .92$. This strongly suggests that the scale scores reflect a common latent variable.

Another obvious difference between the instruments is evident in inspection of the response frequency distributions in Tables 1 and 2. As noted previously, the two instruments use different response formats. The CGI-CogS has a one to seven scale with anchors: none, minimal, mild, moderate, serious, severe, dangerous. However, there were so few ratings in categories six (severe) and seven (dangerous), responses were collapsed in Table 1. SCoRS has a one to four response scale with anchors: none, mild, moderate, and severe. Given the need to collapse the extreme categories with CGI-CogS, and the low rating frequencies in category four on SCoRS, it is apparent that our sample did not contain many participants with extreme (rated severe and above) levels of cognitive deficits. This is understandable given that the sample was intentionally selected to involve schizophrenia out-patients who were on stable medications and not in an acute psychotic episode.

Despite the collapsing of the top two categories, it appears that the CGI-CogS distinction between minimal and mild impairment (versus SCoRS none to mild) reduces the floor effect, whereas at the other end of the scale, the CGI-CogS distinction between moderate and serious (versus SCoRS moderate to severe) better spreads ratings at the high end. Interestingly, judging by the item means and relative frequencies of category ratings, on average participants are rated in the minimal (rating of 2) to mild (rating of 3) impairment range on CGI-CogS items but are characterized as mildly impaired (rating of 2) on the SCoRS items. Such findings may indicate that it is easier for raters to identify impairments using the CGI-CogS instrument relative to the SCoRS. Another way of stating this is that the criterion for a high item rating on SCoRS items may be set too high. Consider for example the SCoRS “forgets names” item cited above where the extreme ratings were: 3) moderate = frequently forgets names, and 4) severe = almost always forgets names. Perhaps “frequently” and “almost always” are too extreme.

Understanding the psychometric differences between the two instruments requires acknowledging their different approaches to measurement. First, as discussed, due to differences in the number of item categories, CGI-CogS items have more response variance relative to SCoRS even after collapsing categories. Although this may partially explain the increased item-intercorrelations for CGI-CogS, the number of response categories cannot completely account for the observed differences. For example, when we collapsed responses on each instrument down to three categories for all items, CGI-CogS still retained a notable advantage in terms of item-test correlations, inter-item correlations, and coefficient alpha. To better understand the differences between the item responses produced from each instrument, we believe that it is important to consider the structure and format of the item content and scoring for the two instruments.

Although both instruments attempt to cover similar aspects of cognitive functioning, CGI-CogS contains item content that explicitly attempts to capture a theory of cognitive deficits that specifies that there are seven “domains” (Nuechterlein et al., 2004) of cognition. Accordingly, CGI-CogS contains two, three, or four items that were written explicitly to capture these domains. In addition, items are administered in order by domain. Because of this explicit domain structure, items within a domain tend to be relatively highly correlated, and thus CGI-CogS has a relatively higher average inter-item correlation. Indeed, as will be demonstrated in the subsequent factor analyses, the CGI-CogS item ratings within a domain do tend to cluster together in a more coherent way relative to SCoRS item ratings.

Another explanation for the difference in internal-consistency is perhaps due to the different way in which the items are written and the content anchoring for rating the responses. As noted previously, each CGI-CogS item is phrased as a series of examples (e.g., CGI-CogS Item #3: “Do you have trouble concentrating? Do you take breaks frequently? Do you have trouble paying attention while reading, listening to the radio, or watching television, long enough to read/listen/see a whole article/chapter/program?”). If the participant or informant acknowledges difficulties, they are then asked to discuss concrete examples which are written down, and further probed by the interviewer. Finally, the content anchors for the response categories request an evaluation of the degree of cognitive deficit and the extent to which that deficit interferes with day-to-day functioning (e.g., 4) moderate = “cognitive deficits with clear effects on functioning”). In contrast, SCoRS items tend to be more specific in content, and the content anchors for the response categories are defined by frequency or time judgments. For example, SCoRS Item 3 asks “do you have difficulty... following a TV show: examples, favorite shows or news.” Then based on what the participant or informant describes, the rater selects: mild = “can follow a short movie or news show (1–2 hours)”, moderate = “can only follow a light 30 minute show (i.e., sitcom)”, or severe = “unable to follow a TV show for any period of time more than five minutes.”

In sum, CGI-CogS items probe about a wide range of contexts where problems (e.g., concentration) may arise. In turn, CGI-CogS items may lead to a more informative interaction among the interviewer and interviewee and provide the clinical interviewer with a more precise view of overall cognitive impairment and its impact on daily functioning. In contrast, SCoRS items tend to ask about a specific context (e.g., watching a movie or TV show) and then the interviewer must judge whether there is an impairment, and whether that impairment occurs rarely, frequently, or nearly all the time. One conclusion we draw from these findings is that although the high correlation between raw scores indicates that CGI-CogS and SCoRS items are measuring essentially the same underlying construct, the CGI-Cogs produces more internally consistent ratings. The consequences of this will become very clear after we review the factor analytic and item response theory results in the following sections.

Dimensionality

In this section, we explore the dimensionality of responses to the CGI-CogS and SCoRS instruments. One goal of this research is to apply an IRT model so that we can make more informed decisions regarding shortening the instruments and to better understand how the items function as a measure of cognitive deficits. Because the IRT model we use makes a “unidimensionality” assumption, we need to explore whether this assumption is reasonably satisfied in these data sets. Beyond these psychometric objectives, exploring the dimensionality of item responses derived from these instruments is also of substantive interest. For example, as noted previously, both CGI-CogS and SCoRS were written to recognize diverse aspects of cognitive functioning. It is interesting to ask, to what degree do these domains emerge as multidimensionality, and in turn, does this multidimensionality interfere with our ability to fit an IRT model and scale individuals on a common dimension?

The first step in the analyses was to estimate a polychoric correlation matrix for each scale separately. Polychorics were estimated using the POLYMAT command available in the PSYCH package (Revelle, 2009) of the *R* statistical software (*R* Development Core Team, 2008). This step was necessary to avoid the well known problems of factor analyzing item-level data using Pearson correlations. For each scale separately, we then conducted three separate sets of exploratory factor analyses: 1) a unidimensional model, 2) a multidimensional correlated factors model, and 3) a bifactor model (Schmid & Leiman, 1957). In the unidimensional model we inspected the eigenvalues, percent of variance explained, mean and standard deviation of residuals, and a goodness-of-fit index (GFI;

McDonald, 2000). The GFI index reflects the proportion of common variance explained by the model. Although there are no gold-standard rules-of-thumb for deciding when a response matrix is “unidimensional enough” for IRT modeling (see Embretson & Reise, 2000), generally researchers look for a large ratio of the first to second eigenvalues (e.g., > 3 to 1), a goodness-of-fit at least 0.90 (better if above 0.95), and residuals with a mean of 0.0 and standard deviation less than 0.05.

We then extracted for each scale from 1 to 7 factors with direct oblimin ($\delta = 3$) rotations. These solutions were evaluated using the same indices as described above, but in addition, we examined the substantive coherence of a factor, and the correlation among the extracted factors. As argued elsewhere (Reise, Morizot, & Hays, 2007), multidimensional solutions are useful in identifying item content clusters in correlation matrices, but they are not very useful in terms of deciding whether the data are unidimensional enough for IRT analyses. As is well recognized (McDonald, 1981), the item responses that result from a measure of a substantively complex construct are rarely if ever strictly unidimensional (i.e., each item has only one *common* cause, the rest of the variation is item specific and error). Typically, many psychological measures contain repeated content which, in turn, will emerge as factors in a factor analyses, and if sample size is large enough, they will be statistically significant (see Gibbons & Hedeker, 1992, for similar arguments). Yet, the presence of multidimensionality does not necessarily vitiate either the scoring of an instrument as a measure of a single construct or the fitting of a subsequent unidimensional IRT model. The important criterion is whether there is a dominant general factor running through the items. The way to explore this issue, as argued by Reise, Morizot, and Hays (2007) and others (e.g., Immekus & Imbrie, 2008), is through fitting a bifactor model, and comparing the results to the unidimensional model.

Because exploratory (as opposed to confirmatory) bifactor analysis may be unfamiliar, we will briefly review the method. Note first that commonly used factor analytic rotation methods are fundamentally incapable of identifying (rotating to) a bifactor structure where each item loads on a common factor, and one or more item content based factors called “group” factors. The reason is that factor rotations (e.g., varimax, promax, and oblimin) that are included in commonly-used software are designed to identify “simple structures” where each item loads only on a single factor (see Browne, 2001). However, this presents a problem for researchers who want to explore alternative factor structures such as a bifactor.

There are three solutions to this dilemma. First, a researcher may use targeted factor rotations as are available in the Comprehensive Exploratory Factor Analysis program (Browne, Cudeck, Tateneni, & Mels, 2004). These rotations are technically complicated and will not be discussed further here. Second, a researcher may avoid the rotation issue by imposing a “confirmatory” bifactor structure and then use structural equations modeling software to fit a model. This topic will also not be discussed because our goal is to explore data structure, not to confirm a model or compare the statistical fit of two alternative models.

A third solution, and the one used here is to conduct an exploratory bifactor analysis using the Schmid-Leiman (1957) orthogonalization technique. To implement the Schmid-Leiman, we used the SCHMID command available in the PSYCH package (Revelle, 2009) of the *R* statistical software. In an exploratory bifactor model, each item is allowed (but not forced) to load on a general factor and one or more “group” factors. The group factors potentially account for residual common variance that is left unexplained by the general factor. Typically, group factors emerge when measures contain clusters of items with highly similar content. In a bifactor model, the general and group factors are constrained to be orthogonal.

Essentially, an exploratory bifactor model (as opposed to confirmatory) is an "orthogonalization" of a second-order factor model. That is, first a set of correlated factors are extracted and rotated (e.g., using principal axis extraction with oblimin rotation). These correlated factors are called "primary" or "first-order" factors. Then, the correlations among those primary factors are factored, and the resulting factor loadings reflect the correlations between the primary and "higher-order" dimension. The Schmid-Leiman technique transforms this higher-order solution into a bifactor structure where each item can load on a common dimension and one or more group factors. Generally speaking, if the original oblique factor extraction contains three correlated factors, then the Schmid-Leiman will contain one general and three group factors. Factor identification and interpretability is judged by looking for an independent cluster basis (i.e., each factor has three or more items that load only on that factor) as explained in McDonald (1999; 2000).

In sum, in an exploratory analysis (but not confirmatory), an oblique rotation, a higher-order factor model, and a Schmid-Leiman orthogonalization are all equivalent ways of examining data and are highly interrelated (see Chen, West, & Sousa, K. H., 2006, and the citations therein for further discussion). Indeed, after specifying the number of primary dimensions (or group factors) the SCHMID command in the PSYCH package produces all three solutions. Finally, we note that the Schmid-Leiman can be easily implemented in SPSS and SAS as well (e.g., Wolff & Preising, 2005).

To evaluate the bifactor model, and ultimately the viability of a unidimensional IRT model, we first evaluate the relative percent of variance explained by the general factor with the percent explained by the group factors. To the degree that the observed variance is explained by a general factor relative to the group factors, a unidimensional IRT model is viable. A second method of evaluation is to compare the factor loadings in the general factor of the bifactor model with the factor loadings in the unidimensional model. If the factor loadings are similar in the unidimensional model (i.e., a model with loadings potentially biased by multidimensionality), and the general factor in the bifactor model (i.e., a model that takes into account multidimensionality), this is further evidence that multidimensionality can safely be ignored and treated as nuisance.

CGI-CogS

The first seven eigenvalues from the CGI-CogS item responses are: 11.56, 1.44, 1.11, 0.93, 0.68, 0.62, and, 0.59, respectively. This dramatic 8 to 1 ratio of the 1st to 2nd eigenvalue suggests a large general factor. Under the Uni column of Table 3 are the loadings for the unidimensional model, and it is clear that each item loads strongly on this first dimension with loadings ranging from .57 to .81. The GFI index, which reflects the degree to which common variance is accounted for is .98, strongly supporting the adequacy of a single factor. The mean residual is zero and the standard deviation of the residuals is .07. This latter value is a little above our preferred benchmark, but still it is in an acceptable range.

Despite this strong evidence for a unidimensional solution, for the sake of thoroughness, we also explored multidimensional solutions. Although solutions of five, six, and seven correlated dimensions were evaluated, these results are difficult to interpret due to Heywood cases, factors defined by single items, or no item having its highest loading on a factor. Thus, we argue that the only viable multidimensional solution is the four factor, and it is this model that is reported in the middle set of columns in Table 3. In this model, GFI is .99 the mean residual is .03 and the standard deviation of the residuals is .03. The factors are clearly content domain factors where Factor 1 is merged attention and reasoning domain items, Factor 2 is the four social cognition domain items, Factor 3 is spatial and verbal learning domain items, and Factor 4 consists of items from the speed of processing domain. The

factors are strongly inter-related with correlations ranging from .52 (3 and 4) to .69 (1 and 2).

The four factor model is an “improvement” on the unidimensional model in terms of GFI and standard deviation of residuals. However, it is not surprising that a model with more factors explains the item response data better. To explore the extent to which this is a meaningful improvement, we turn to the bifactor results, which are reported in the third set of columns in Table 3. Given the exploratory factor analysis results, we report on only the findings of the one general factor and four group factor solutions. In the exploratory bifactor model ($GFI = .99$), all items continue to load highly on the general factor but the loadings are around .03 to .05 smaller in the general factor than in the unidimensional model. This suggests that the multidimensionality “biases” the loadings in the unidimensional solution, but only slightly. The group factors replicate the four factor model above, with the exception that now the loadings are much smaller due to the fact that the general factor has been controlled for in the bifactor solution. Finally, we note that in the bifactor solution, the general factor accounts for 47% of the variance (73% of the common variance), whereas the group factors explain, 4%, 4%, 4%, and 5%, respectively, or 17% combined.

SCoRS

The first seven eigenvalues from the SCoRS were: 7.86, 1.45, 1.36, 1.15, 0.99, 0.85, and, 0.74, respectively, a pattern suggesting the presence of only a single meaningful dimension. The ratio of the 1st to 2nd eigenvalue is 5.42, which is less than CGI-CogS, but still well above the standard 3 to 1 ratio used to judge unidimensionality. In the first set of factor loadings in Table 4 are the loadings for the unidimensional model, and again, each item loads strongly on this single dimension with loadings ranging from .49 to .79. The GFI index is .97, strongly supporting the adequacy of a single factor. The mean residual is 0.0 and the standard deviation of the residuals is .08. As before with CGI-CogS, this latter value is above our preferred benchmark, but still it is by no means a large value.

We also explored multidimensional solutions with SCoRS ranging from 1 to 7 correlated dimensions extracted. However, the interpretation of the dimensions was more challenging in SCoRS relative to the CGI-CogS. In addition, we again had concerns with factor solutions above four dimensions, namely, single item factors, factors having no item with the highest loading, or factors that were impossible to meaningfully interpret. For these reasons, we report the results of the four correlated factor solution in the middle columns of Table 4. In this model, GFI is .99, the mean residual is 0.0 and the standard deviation of the residuals is .04. We interpret the factors as follows. Factor 1 appears composed of attention and focus items. Factor 2 is a memory factor containing four items with the word “remember” in it and “learns new things.” Factor 3 has item content reflecting understanding of meaning and is similar to CGI-CogS social cognition. Factor 4 is marked by two items, “familiar tasks”, and “remembers chores,” which is perhaps a demands of daily living factor. The SCoRS factors are not as highly inter-related with correlations ranging from 0.34 (4 and 2) to 0.54 (1 and 2).

In the last set of factor loadings in Table 4, we report on the findings of the one general factor and four group factor solutions. In this exploratory bifactor model ($GFI = .99$), all items continue to load highly on the general factor but the loadings are around .03 to .06 smaller in the general factor than in the unidimensional model. As expected, the factor loadings on the group factors are now smaller due to the fact that the general factor has been controlled for in the bifactor solution. The general factor accounts for 30% of the item response variance (60% of common variance), whereas the group factors explain, 6%, 5%, 4%, and 5%, respectively, or 20% combined. Clearly, the general factor is not as strong in

SCoRS relative to CGI-CogS, and the factors do not cohere as one would expect based on the *a priori* content groupings.

Combined

Although the above analyses were important for us to learn how each instrument functions as a stand alone measure, it is these combined analyses that are most decisive in terms of evaluating the tenability of whether to fit an IRT model to both scales simultaneously. The first seven eigenvalues of the combined matrix are: 18.22, 2.17, 1.75, 1.52, 1.51, 1.40, and 1.28, respectively, again suggesting a strong general factor. To save space, and because the multidimensional solutions were so challenging to interpret, we will not provide a table of factor loadings (available from authors) or provide a detailed description of the multidimensional solutions in this section. Rather, we note first that in the unidimensional solution, all items continued to load substantially. Second, although solutions containing from four to seven group factors were considered in the bifactor analysis, we decided to extract seven group factors. The GFI index for this model as 0.99 with a mean residual of 0.00 and standard deviation of 0.03.

As with the previous results, in bifactor solutions, the loadings on the general factor were only slightly lower in the bifactor relative to their values in the unidimensional solution. This suggests that the multidimensionality does not interfere with our ability to use the items to scale individuals on a common dimension. Third, we note that unsurprisingly given the above results, with the exception of SCoRS item #14 (“learns new things”) and #20 (“follows group conversations”), the CGI-CogS items dominate the general factor in terms of loadings. Finally, it is clear that when the two instruments are combined, the general factor continues to dominate the group factors. Specifically, the general factor accounts for 38% of the variance while the group factors combined account for 23%. However, the dominance of the general factor is reduced in this joint analysis compared to when the CGI-CogS was analyzed alone. Nevertheless, we argue that these findings support our application of IRT to the combined measures, a topic we now address.

Item Response Theory Analysis

Item response theory (IRT; Embretson & Reise, 2000) is the dominant measurement theory in aptitude, achievement, and licensure testing and is quickly becoming the default psychometric method in patient reported health outcomes measurement as well (see Reise & Waller, 2009). IRT measurement models are a device to study the relation between individual differences on a latent variable (e.g., cognitive deficits) assumed to underlie item responses and the probability of responding in a particular response category. IRT measurement models require that certain assumptions be met in order for the results to be interpretable. First and foremost, unidimensional IRT models (i.e., models with only a single latent variable) require that the data be reasonably unidimensional; that is, one common trait should account for the item responses. Another way of stating this is that IRT models assume that the estimated item parameters reflect the relation between item responses and the common trait they are assessing and are not overly distorted by multidimensionality in the data. The above factor analyses establish that although both instruments contain some multidimensionality due to content clusters, the overwhelming majority of common variance is explained by a strong general factor. For this reason, we feel comfortable moving forward with fitting an IRT model.

The Graded Response Model

The graded response model (GRM; Samejima, 1969) is appropriate when items contain ordered categorical responses (Embretson & Reise, 2000; Ostini & Nering, 2006). In the GRM, each scale item (i) is described by one item slope parameter (α_i) and $j = 1 \dots m_i$

“threshold” parameters (β_{ij}), where m is the number of response categories minus one. These parameter estimates allow researchers to compute category response curves (CRCs) that display the relation between a latent variable (cognitive deficits) and the probability of responding in each category. For example, Figure 1 displays the CRCs for SCoRS item #13 “*Stays Focused*”. This item will be discussed in more detail shortly.

As for interpreting the parameters, the β_{ij} parameters in the GRM represent the trait level necessary to respond above a between category boundary j with .50 probability. In other words, for a four-category item scored as 1 through 4, there are three boundaries between the categories (e.g., 1 versus 2, 3, 4; 1, 2 versus 3, 4; and 1, 2, 3, versus 4). Thus, the first threshold parameter represents the point on the latent trait continuum where the probability of responding above category 1 is 50%; the second threshold parameter represents the point on the latent trait continuum where the probability of responding above category 2 is 50%; and the third threshold parameter represents the point on the latent trait continuum where the probability of responding above category 3 is 50%. The slope parameters (α_i) in the GRM are analogous to factor loadings, and are related to the discriminating power of an item. Items with relatively high slope parameters are better able to differentiate among people at different ranges of the latent trait.

In this research, we estimated the GRM parameters for the combined pool of 41 items (21 CGI-CogS and 20 SCoRS items) using MULTILOG (Thissen, 1991). Mostly program defaults were used with two important exceptions. First, the number of estimation cycles allowed was set to 500. This proved to be more than enough iterations to achieve a converged solution. Second, the quadrature points used to define the distribution of the latent trait in the population were specified to range from -4.5 to 4.5 in increments of 0.1 . This was implemented in order to better estimate extreme threshold parameters (e.g., greater than 4.0 or less than -4.0) that may result from low response rates in the extreme categories.

Note that in IRT modeling, for identification purposes, it is customary to define the latent variable (cognitive deficits) such that the mean is zero and the standard deviation is 1.0 . Thus, the item parameter estimates in this study should be interpreted in regard to this metric. Specifically, the mean of the latent trait distribution in this study refers to the mean level of cognitive deficits in a stabilized out-patient population characterized by mild to moderate levels of cognitive impairment.

Finally, due to the fact that for many items, especially SCoRS items, there are few responses in the extreme impaired category, we were concerned with our ability to estimate some threshold parameters. To address this concern, we ran MULTILOG in an iterative fashion. First, we ran the program specifying that all the CGI-CogS items had five categories, and all the SCoRS items had four categories. We then inspected the threshold parameters and their standard errors to check whether the values and standard errors were reasonable. If any estimated threshold parameter was above 4.0 or below -4.0 , we collapsed categories and re-estimated the model. In the first run, we found that the fourth threshold parameter for CGI-CogS #17 was above 4.0 and had a relatively large standard error. In addition, this also occurred for the third threshold parameter for nine of the twenty SCoRS items. We thus collapsed responses in category five into category four for CGI-CogS #17, and collapsed responses in category four into category three for nine SCoRS items. The need to collapse categories reflects both the lack of severely impaired participants, as well as the extremity of the anchor for the extreme category (e.g., forgets names “all the time”).

After collapsing categories, the second MULTILOG run produced more reasonable and interpretable estimates (i.e., threshold estimates between -4 and 4) for all parameters. These parameter estimates along with their standard errors are reported in Table 5. In the first

column of Table 5 are the estimated item slope parameters. The size of the item slope parameters ranges from around 1.0 to 2.5. It is clear that the CGI-CogS items tend to be more discriminating (i.e., better able to differentiate among people) relative to the SCoRS items. This result is completely consistent with the higher item-test correlations and factor loadings in the unidimensional solution for CGI-CogS relative to SCoRS reported earlier. Within CGI-Cogs, all items have reasonably high slopes and there does not appear to be any consistent link between content domain and the size of the estimated slope parameter.

As noted, relative to CGI-CogS, the SCoRS items tend to have the relatively lower slope estimates. This remains true even for SCoRS items that did not contain collapsed categories. One reason this may occur is that the response format for rating CGI-CogS items does a better job of differentiating among people than the response format for rating SCoRS items. In particular, observe that the SCoRS items with the lowest slopes are also the ones with low frequencies in the top category. What happens with these types of items is that responses get bunched up in the lower response categories, which is another way of saying the response categories are not as relatively discriminating. For example, consider the category response curves for SCoRS item #13 “*Stays Focused*” shown in Figure 1. This item has slope of 0.96 with thresholds of -1.99, 1.39, and 3.48, the latter value indicating that the model predicts that an individual must be over three standard deviations above the mean to have a 50% chance of being rated in category four. For a very wide range of the latent trait, a rating in category two is most likely. It is only very extreme individuals (above roughly 1.5 standard deviations) who are more likely to be rated a three, and only the most extremely impaired (above roughly 3.5 standard deviations from the mean) are likely to receive the four rating.

The estimated threshold parameters and their standard errors are shown in the next set of columns. Recall that these values reflect the point on the latent trait continuum where the probability of responding above a between category boundary is 0.5. For example, using the 1, 2, 3, 4, 5 response scale of CGI-CogS Item #1, the thresholds reflect the trait level necessary to be rated above category 1 (-1.93), above category 2 (-0.57), above category 3 (0.60), and above category 4 (1.85). These location parameters are critically important for a number of reasons. As explained below, they determine where on the latent trait continuum an item provides information among individuals (i.e., where on the latent trait the item discriminates best among individuals). They are also important for judging the item quality and determining how the clinical raters are using the categories to differentiate among people.

Item and Test Information and Standard Error of Measurement

An important difference between traditional measurement theory and IRT is that in the former, individuals are assumed to have equal standard errors regardless of their position on the construct. In IRT, individuals’ can have different standard errors depending on how discriminating or informative a set of items are in different range of the latent trait (cognitive deficits). Specifically, the estimated item parameters from the graded response model (Table 5) can be transformed into item information curves (Ostini & Nering, 2006). Items with larger slope parameters provide relatively more information, and where that information is located on the latent trait continuum depends on the item threshold parameters; item information is maximized around the threshold parameters.

To illustrate, Figure 2 provides the contrast in information between CGI-CogS #14 (slope = 2.41; “judgment in situations”) and SCoRS #9 (slope = 1.34; “keep track of money”). Because information is a function of the square of the item slope, the CGI-CogS item has nearly three times the information in the middle of the trait range. In short, it would take three items like SCoRS #9 to discriminate as well as one CGI-CogS #14. For a second example, in Figure 3 we provide a contrast of the information curves for CGI-CogS #4

(“focus on information”) and CGI-CogS #11 (“initiate attempts to solve problems”). These two items have similar slopes, but somewhat different threshold parameters. Although the differences in item information are not dramatic, the former item provides relatively more information in the high range of cognitive deficit whereas the latter provides relatively more in the low range.

Information is a critically important concept in IRT for three reasons. First, information is additive across the items administered. Thus, if a researcher wanted to know how informative a scale consisting of the first five CGI-CogS items would be, all he or she would have to do is add together the item information functions for those items. For example, in Figure 4 is shown the scale information for all 41 items, the 21 CGI-CogS items and the 20 SCoRS items. Taken as a whole or by scale, this pool of items is most informative around the mean trait level and then tapers off sharply at both extremes. Second, information is inversely related to an individual’s standard error of measurement; the more information a measure provides in a certain trait range, the more precise an individual’s measurement in that trait range will be. This feature of IRT allows researchers to go beyond asking how reliable or precise a score is, and to study how precise a measurement is for individuals in high, medium, and low trait ranges. A third important feature of information, as demonstrated below, is that it plays a crucial role in real-world or simulated computerized adaptive testing (CAT; Wainer et al., 2000). Specifically, in an adaptive test, the item information function is used to select items that are most informative (i.e., discriminating) for a particular individual dependent on their currently estimated trait level as they progress through a computerized test. In the following section, we explore the use of simulated adaptive testing in order to further learn about the relative quality of the items and to explore the consequences of shortening the combined CGI-CogS and SCoRS measure.

Simulated Computerized Adaptive Testing: How Many Items are Really Needed?

In traditional psychometrics, to compare individuals on the same scale, all people must be administered the same items or parallel tests. In contrast, under an IRT framework, an individual’s position on the latent trait scale can be estimated based on any subset of items with known item parameters (Embretson & Reise, 2000). For this reason, it is often argued that one of the major advantages of IRT measurement models is that they form the basis for computerized adaptive testing. Indeed, computerized adaptive test administration has been used extensively for large-scale aptitude and licensure testing (see, Wainer et al., 2000) and is increasing popular in personality and health outcomes assessment (Reise & Waller, 2009).

In this study, we use the logic of IRT-based CAT to explore the question, if hypothetically a CAT were to be administered, which items would be selected the most, and in turn, how many items are needed to scale individuals on a cognitive deficit continuum at different levels of precision? We explored these issues by performing a computerized adaptive simulation using the estimated item parameters from the previous section, and the actual item response patterns from the 315 data points with complete responses to all items. This simulation is easily accomplished using an existing software package called FIRESTAR 0.09 (Choi, 2009). This program reads in the estimated item parameters and then computes item information functions as described in the previous section. We specified that all individuals should start at a latent trait value of 0.0. For each individual response vector, FIRESTAR then finds the item that is most informative at latent trait equals zero (note: this is always the same item for all individuals). FIRESTAR then finds the response and re-estimates an individual’s trait level and standard error using the expected a posteriori (Mislevy & Bock, 1982) method. This cycle continues until a stopping rule is encountered. In this study, we explored four different stopping criteria. Specifically, we instructed FIRESTAR to end adaptive testing when a person’s standard error was either at or below: .

25, .3, .4, or .5. On a z-score metric, these values are roughly analogous to error variances of 6%, 9%, 16%, and 25%, respectively.

Real-data simulations are typically judged by several criteria. First, how many items were needed to reach the stopping criteria? In these simulations, the average number of items administered was 14.36, 8.68, 4.39, and 2.63, for the four standard error criteria, respectively. A second criterion addresses the question: how does the number of items needed change as a function of position on the latent trait? In Figure 5, the x-axis is the latent trait score estimate based on all 41 items and the y-axis is the number of items administered during simulated CAT for the 0.30 standard error condition (note: other conditions displayed similar patterns). This figure demonstrates that for people in the middle of the cognitive deficits trait range (-2 to 2), very few items are needed due to the fact that there are many very informative items in that trait range. However, for individuals at the very low end of cognitive deficits (< -2.0), even the entire pool may be unsatisfactory – the present instrument simply does not allow for high measurement precision at the low range. It is arguable that this is irrelevant for this construct because low levels of cognitive deficits would not be a target for intervention.

A third way to evaluate a real-data simulation is to examine the correlation between latent trait scores estimated based on CAT versus those based on administration of all 41 items. In the present case, the more stringent the stopping criterion, the higher the correlation. Specifically, the observed correlations are $r = .98, .96, .92$, and $.89$, respectively for the four conditions. Of course, these values are inflated due to shared items in the CAT and in the full-scale trait level estimate. Given this, it is still remarkable that the full-scale score can be reproduced almost perfectly with a little over one-third the item pool (i.e., around 14 items).

Finally, it is important to inspect the frequency distribution of which items were administered. For the $SEM < 0.3$ condition, for example, most examinees receive the same 8 to 12 items. Specifically, CGI-CogS items 1, 2, 4, 6, 11, 14, and to a lesser extent, items 12, 3, 19, and 15 were administered to most individuals regardless of their trait standing. In Figure 4 we include the scale information function based on these 10 items. Interestingly, these 10 CGI-CogS items provide more information than the 21 SCoRS items. Given that a SEM of less than 0.3 implies an error variance of 0.09 (reliability 0.91), it is clear that in terms of scaling individual differences, the CGI-CogS and SCoRS combined item pool can be seriously reduced while still maintaining reasonably precise latent trait score estimates. Such results suggest that future researchers consider either a fixed-length short-form or a computerized adaptive test. However, given our findings that even in a CAT, the same set of 10 items are consistently administered to most people, it would be difficult to argue for a CAT over a fixed-length short form in this context.

Discussion

A main objective of this research was to thoroughly explore the psychometric properties of two instruments designed to measure cognitive functioning in schizophrenia patients – CGI-CogS (Ventura et al., 2008) and SCoRS (Keefe et al., 2006). These measures were developed to primarily serve as co-primary measures in clinical trials of new pharmaceutical treatments for cognitive impairments in individuals with schizophrenia and other disorders. Other unexplored uses of these instruments are as measures of change in cognitive functioning and as screening tools. We believe that our results are informative regarding all these potential applications.

Although the instruments are similar in content and raw scores on the two instruments are highly correlated, item ratings on CGI-CogS tended to have superior psychometric

properties, relative to SCoRS. Specifically, item ratings to CGI-CogS displayed higher average item-intercorrelations, higher item-test correlations, and thus, higher internal consistency reliability estimates. In turn, CGI-CogS items had relatively larger factor loadings in a unidimensional solution, higher slope parameter estimates in application of IRT, provided more information across the trait range, and were always the first couple of items administered in simulated computerized adaptive testing.

We do not attribute these differences merely to differences in the number of response categories. Rather, as described in a previous section, we attribute these differences to: a) the differential use of prompts (i.e., CGI-CogS provides many prompts to begin an item), b) how the items are phrased (i.e., CGI-CogS examines functioning across multiple contents whereas SCoRS items examine functioning in one area at a time), and c) how the items are rated (i.e., CGI-CogS items are rated on general impairment in functioning while SCoRS items are rated on the relative frequency of impairment in a specific domain of functioning).

Our analyses of category response frequencies and IRT threshold parameters also revealed some interesting differences between the data derived from two interview-based rating instruments. As noted, CGI-CogS has a seven-point format while SCoRS uses a four-point. For both instruments, ratings in the highest categories were sparse. This is attributable partly to our sample, which may not include schizophrenia patients with the most severe cognitive impairment due to the fact that the sample was selected to be typical of outpatient samples in clinical trials of new cognitive-enhancing interventions (Nuechterlein et al., 2008). Nevertheless, even after reducing the CGI-CogS response options from seven to five, the CGI-CogS appeared to do a relatively better job of limiting floor effects. Also, inspection of Table 5 reveals that the IRT threshold parameters for CGI-CogS are better spread out over a reasonable range of the trait (i.e., from -3 to $+3$). The threshold parameters for SCoRS item for the higher rating points are routinely large, and almost always higher than comparable values for CGI-CogS items. Certainly more research is needed into the optimal number of response options and anchor labels. Too many response options forces raters to make distinctions that may not produce reliable scores, whereas too few can miss important distinctions.

Our next set of analyses were three types of factor analyses – unidimensional, multidimensional, and bifactor. The factor analytic results were compelling in that a strong general factor dominates the common variance among items of these interview-based measures of cognitive functioning. Although within each instrument significant multidimensionality was identified, multidimensional models showed that the dimensions are highly correlated, especially on CGI-CogS. More importantly, bifactor analyses revealed that: 1) multidimensionality caused by item content factors does not vitiate the ability of the scales to measure a common dimension (i.e., loadings on the general trait in the bifactor were high and not substantially different in value from the loadings in the unidimensional model), and 2) if subscales were to be formed based on content factors, scores on those scales would have nearly zero reliability after the affect of the general factor was controlled. In short, despite multidimensionality caused by heterogeneous item content, the only reliable rating that can be assessed with these two instruments is a common deficit factor. This is true even for the CGI-CogS, which was designed to be content heterogeneous and multidimensional in the hope that differential change in different domains of cognition could be measured. Thus, interview-based ratings of cognitive functioning may reflect a single dimension of cognitive deficit to a greater extent than the objective cognitive performance measures that have been the data source for most prior factor analytic studies of cognition in schizophrenia (Nuechterlein et al., 2004). Further research comparing interview-based and performance-based measures of cognition in schizophrenia will be needed to shed more light on this issue.

Our final set of results revolved around using IRT modeling and simulated adaptive testing to explore the extent to which the scales could be reduced. Given that the items had high discrimination, which results in higher levels of item information, it was not surprising that our simulated adaptive testing results showed that “in theory” individuals can be well assessed using a much smaller pool of items. Specifically, simulated adaptive testing indicated that administering between 8 and 12 items resulted in trait level estimates that were highly correlated with trait level estimates derived from the entire pool of 41 items. This suggests that interview-based measures of cognitive deficits could be greatly shortened without significant loss of measurement precision. In fact, by considering both content representativeness and item-level psychometric properties reported here, we have recently developed a 10-item short-form version, the Cognitive Assessment Interview or CAI (Ventura, et al., under review). Initial findings suggest that scores derived from the 10-item version have very similar patterns of external correlations as the entire 41-item pool. It is hoped that further investigation of these new instruments will help advance treatment research in schizophrenia and other mental disorders, and that their use will provide new insights about the meaning of cognitive impairment to both patients and their caregivers.

Acknowledgments

Funding for the MATRICS Initiative was provided through an NIMH contract N01MH22006 to UCLA (PI: Steven Marder, M.D., Co-PI: Michael Green, Ph.D.). This research was also supported by an investigator initiated grant from Pfizer, Inc. (PI: Joseph Ventura, Ph.D.) and from an NIMH R21 grant 1R21MH073971 (PI: Joseph Ventura, Ph.D.). Additional funding support was provided by MH37705 (PI: Keith H. Nuechterlein, Ph.D.) and P50 MH066286 (PI: Keith H. Nuechterlein, Ph.D.). This work was also supported by the Consortium for Neuropsychiatric Phenomics (NIH Roadmap for Medical Research grants UL1-DE019580 (PI: Robert Bilder), and RL1DA024853 (PI: Edythe London).

References

- Browne MW. An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*. 2001; 35:111–150.
- Browne, MW.; Cudeck, R.; Tateneni, K.; Mels, G. CEFA: Comprehensive Exploratory Factor Analysis. 2004. Version 2.00 [Computer software and manual]. Available at <http://quantrm2.psy.ohio-state.edu/browne/>
- Buchanan RW, David M, Goff D, Green MF, Keefe RSE, et al. A summary of the FDA-NIMH-MATRICS workshop on clinical trial design for neurocognitive drugs for schizophrenia. *Schizophrenia Bulletin*. 2005; 31:5–19. [PubMed: 15888422]
- Chen FF, West SG, Sousa KH. A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*. 2006; 41:189–225.
- Choi SW. Firestar: Computerized adaptive testing (CAT) simulation program for polytomous IRT models. *Applied Psychological Measurement*. 2009; 33:644–645. [PubMed: 20011609]
- Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Erlbaum; 2000.
- Gibbons RD, Hedeker D. Full-information item bifactor analysis. *Psychometrika*. 1992; 57:423–436.
- Green M. Cognition, drug treatment, and functional outcome in schizophrenia: a tale of two transitions. *American Journal of Psychiatry*. 2007; 164:992. [PubMed: 17606645]
- Green MF, Nuechterlein KH. The MATRICS initiative: Developing a consensus cognitive battery for clinical trials. *Schizophrenia Research*. 2004; 72:1–3. 1. [PubMed: 15531401]
- Green MF, Nuechterlein KH, Kern RS, Baade LE, Fenton WS, Gold JM, Keefe RS, Mesholam-Gately R, Seidman LJ, Stover E, Marder SR. Functional co-primary measures for clinical trials in schizophrenia: Results from the MATRICS Psychometric and Standardization Study. *American Journal of Psychiatry*. 2008; 165:221–228. [PubMed: 18172017]
- Hyman S, Fenton W. What are the right targets for psychopharmacology? *Science*. 2003; 299:350–351. [PubMed: 12532001]

- Immekus JC, Imbrie PK. Dimensionality assessment using the full-information item bifactor analysis for graded response data. *Educational and Psychological Measurement*. 2008; 68:695–709.
- Keefe R, Poe M, Walker T, Kang J, Harvey P. The schizophrenia cognition rating scale: an interview-based assessment and its relationship to cognition, real-world functioning, and functional capacity. *American Journal of Psychiatry*. 2006; 163:426–432. [PubMed: 16513863]
- McDonald RP. The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*. 1981; 34:100–117.
- McDonald, RP. *Test theory: A unified Approach*. Mahwah, NJ: Erlbaum; 1999.
- McDonald RP. A basis for multidimensional item response theory. *Applied Psychological Measurement*. 2000; 24:99–114.
- Mislevy RJ, Bock DR. Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*. 1982; 6:431–444.
- Nuechterlein KH, Barch DM, Gold JM, Goldberg TE, Green MF, Heaton RK. Identification of separable cognitive factors in schizophrenia. *Schizophrenia Research*. 2004; 72:29–39. [PubMed: 15531405]
- Nuechterlein KH, Green MF, Kern RS, Baade LE, Barch DM, Cohen JD, Essock S, Fenton WS, Frese FJ3rd, Gold JM, Goldberg T, Heaton RK, Keefe RS, Kraemer H, Mesholam-Gately R, Seidman LJ, Stover E, Weinberger DR, Young AS, Zalcman S, Marder SR. The MATRICS Consensus Cognitive Battery, part 1: Test selection, reliability, and validity. *American Journal of Psychiatry*. 2008; 165:203–213. [PubMed: 18172019]
- Ostini, R.; Nering, ML. *Polytomous item response theory models*. Thousand Oaks, CA: Sage; 2006.
- R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria: 2008. ISBN 3-900051-07-0, URL <http://www.R-project.org>
- Reise SP, Waller NG. Item response theory and clinical measurement. *Annual Review of Clinical Psychology*. 2009; 5:27–48.
- Reise SP, Morizot J, Hays RD. The role of the bi-factor model in resolving dimensionality issues in health outcomes measures. *Medical Care*. 2007; 16:19–31.
- Revelle, W. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.0–68. 2009. <http://personality-project.org/r/psych.manual.pdf>
- Samejima F. Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Suppl*. 1969:17.
- Schmid J, Lieman JM. The development of hierarchical factor solutions. *Psychometrika*. 1957; 22:53–61.
- Thissen, D. *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software; 1991. (Version 6.0) [Software manual]
- Ventura J, Green MF, Shaner A, Liberman RP. Training and quality assurance on the Brief Psychiatric Rating Scale: the "drift busters". *International Journal of Methods in Psychiatric Research*. 1993; 3:221–224.
- Ventura J, Cienfuegos A, Boxer O, Bilder R. Clinical global impression of cognition in schizophrenia (CGI-Cogs): reliability and validity of a co-primary measure of cognition. *Schizophrenia Research*. 2008; 106:59–69. [PubMed: 17900866]
- Ventura J, Reise SP, Keefe R, Baade LE, Gold JE, Green MF, et al. The cognitive assessment interview (CAI): Validation of an empirically derived, brief interview-based measure of cognition. (under review).
- Wainer, H.; Dorans, NJ.; Eignor, D.; Flaugher, R.; Green, BF., et al. *Computerized adaptive testing: A primer* 2nd Edition. Mahwah, NJ: Erlbaum; 2000.
- Wolff H, Preising K. Exploring item and higher-order factor structure with the Schmid-Leiman solution: Syntax codes for SPSS and SAS. *Behavior Research Methods*. 2005; 37:45–58.

Figure 1.
Category Response Curves: SCoRS #13 Stay Focused.

Figure 2.
Item Information Functions CGI-CogS #14 Versus SCoRS #9.

Figure 3.
Item Information Functions CGI-CogS #11 Versus CGI-CogS #4.

Figure 4.
Scale Information Functions for Entire Pool, CGI-CogS, SCoRS, and 10 CAT Items.

Figure 5.
Plot of CAT Based Trait Level Estimates Versus Number of Items Administered to Reach a Standard Error of Measurement Less Than 0.30 Criterion.

Table 1

Abbreviated Item Content and Descriptive Statistics For the CGI-CogS Scale

Item#	Cognitive Impairment ^a	Mean	Std. Dev.	Item-test	Response Proportions				
					1	2	3	4	5 ^b
CGI 1	Newly Learned Information	2.94	1.01	.71	6.0	28.8	37.3	20.7	7.2
CGI 2	On the Spot Manipulation	2.75	1.24	.71	21.0	20.4	30.1	19.4	9.1
CGI 3	Sustained Concentration	3.07	1.10	.66	10.3	17.9	34.5	29.2	8.2
CGI 4	Focus on Information	2.34	1.14	.74	27.6	33.2	20.7	14.4	4.1
CGI 5	Ignore Distracting Events	3.11	1.11	.61	9.4	17.9	36.1	26.0	10.7
CGI 6	Learn / Remember Verbal Material	3.10	1.09	.69	8.5	18.8	37.6	34.5	10.7
CGI 7	Recall Recent Events	2.24	1.09	.64	30.4	32.9	21.9	11.9	2.8
CGI 8	Remembering Material Seen	1.97	1.00	.64	40.4	31.7	20.1	6.3	1.6
CGI 9	Learning New Routes	2.40	1.19	.57	27.6	30.4	21.9	14.1	6.0
CGI 10	Recalling Locations	2.27	1.17	.60	32.9	29.2	20.4	13.5	4.1
CGI 11	Initiate Attempts to Solve Problems	3.12	1.18	.72	11.3	17.2	32.9	25.4	13.2
CGI 12	Lack of Flexibility	2.85	1.17	.70	16.6	20.1	33.5	21.6	1.8
CGI 13	Lack of Persistence	2.64	1.23	.65	22.6	24.8	27.3	16.9	8.5
CGI 14	Judgment in Social Situations	2.64	1.23	.77	19.7	31.3	24.5	14.4	10.0
CGI 15	Performs Tasks Slowly	2.56	1.23	.68	27.0	21.6	25.7	20.1	5.6
CGI 16	Slow Conversation	2.08	1.14	.62	40.1	29.2	16.6	10.7	3.4
CGI 17	Moves Slowly	1.81	1.00	.50	51.1	26.6	14.4	6.3	1.6
CGI 18	Affect Perception	2.55	1.27	.67	29.2	19.4	25.1	19.7	6.6
CGI 19	Appreciate Intentions	2.83	1.18	.72	15.7	23.5	32.6	18.5	9.7
CGI 20	Appreciate Humor	2.57	1.26	.63	26.3	23.8	24.1	18.2	7.5
CGI 21	Reciprocal Communication	2.32	1.20	.66	33.5	23.8	24.5	13.5	4.7

Note. N = 319. The response categories are anchored as 1 = no impairment, 2 = minimal impairment, 3 = mild, 4 = moderate, and 5 = serious impairment.

^a Cognitive impairment refers to the specific cognitive deficit the item was designed to assess.

^b The original number of response categories was seven. However, due to a lack of ratings in the highest categories, ratings in categories six (severe) and seven (severe/dangerous) were collapsed and rated as a five (serious).

Table 2

Item Content and Descriptive Statistics For the SCoRS Scale

Item#	Cognitive Impairment ^a	Mean	Std. Dev.	Item-test	Response Proportions			
					1	2	3	4
SCR 1	Remember names of people	2.02	.71	.46	22.3	55.4	20.1	2.2
SCR 2	Remember how to get places	1.54	.72	.52	58.2	30.3	10.2	1.2
SCR 3	Follow a TV show	2.10	.84	.50	26.9	39.9	29.1	4.0
SCR 4	Remember where things are put	1.83	.76	.48	37.5	43.3	17.6	1.5
SCR 5	Remember chores and responsibilities	1.79	.74	.49	38.1	47.4	12.1	2.5
SCR 6	Learn how to use gadgets	2.03	.81	.52	30.3	37.5	31.0	1.2
SCR 7	Remember information or instructions	2.18	.74	.60	16.7	52.0	27.9	3.4
SCR 8	Remember what to say	1.83	.70	.55	33.4	51.7	13.6	1.2
SCR 9	Keep track of money	2.12	.89	.50	30.0	32.5	32.8	4.6
SCR 10	Keep from jumbling words	1.56	.65	.44	52.6	39.3	7.4	0.6
SCR 11	Concentrate enough to read the paper	2.34	.81	.53	17.6	35.0	43.3	4.0
SCR 12	Perform familiar tasks	1.38	.65	.42	70.9	21.7	6.2	1.2
SCR 13	Stay focused	2.10	.73	.47	17.6	59.4	18.3	4.6
SCR 14	Learn new things	1.97	.76	.67	30.0	44.0	25.4	0.6
SCR 15	Ideas into words	1.82	.81	.48	39.9	41.8	14.6	3.7
SCR 16	Do things quickly	1.72	.72	.43	43.3	43.0	12.4	1.2
SCR 17	Handle changes in daily routine	1.94	.73	.55	30.0	46.4	23.2	0.3
SCR 18	Understand meanings	1.73	.78	.57	45.2	39.9	11.8	3.1
SCR 19	Understand other's feelings	2.05	.88	.50	31.3	38.1	25.4	5.3
SCR 20	Follow group conversation	1.96	.74	.43	27.2	52.3	17.6	2.8

Note. N = 323.

^a Cognitive impairment refers to the specific cognitive deficit the item was designed to assess. The response categories are rated 1 = no impairment, 2 = mild, 3 = moderate, and 4 = severe impairment.

Table 3

Unidimensional, Multidimensional, and Bifactor Analyses of CGI-CogS

Item	Cognitive Domain	Uni.	Four-Factor Oblique Rotation				Schmid-Leiman Bifactor				
			F1	F2	F3	F4	Gen	GR1	GR2	GR3	GR4
CGI 1	Working Memory	0.76	0.08	0.11	0.64	0.10	0.71	-0.01	0.05	0.05	0.42
CGI 2	Working Memory	0.77	0.62	0.07	0.19	-0.01	0.72	0.32	0.02	-0.03	0.10
CGI 3	Attention	0.72	0.61	-0.06	0.27	0.00	0.67	0.30	-0.07	-0.01	0.16
CGI 4	Attention	0.80	0.69	-0.10	0.35	-0.01	0.75	0.34	-0.09	-0.03	0.20
CGI 5	Attention	0.67	0.42	0.02	0.28	0.03	0.63	0.20	-0.01	0.01	0.17
CGI 6	Verbal Learning	0.75	0.21	-0.02	0.63	0.09	0.70	0.05	-0.04	0.04	0.41
CGI 7	Verbal Learning	0.71	0.17	0.09	0.62	-0.04	0.66	0.04	0.04	-0.06	0.41
CGI 8	Spatial Learning	0.72	0.22	0.04	0.45	0.15	0.68	0.08	0.01	0.08	0.29
CGI 9	Spatial Learning	0.63	0.16	-0.02	0.74	-0.12	0.59	0.02	-0.03	-0.12	0.48
CGI 10	Spatial Learning	0.66	0.03	0.17	0.60	-0.01	0.61	-0.03	0.09	-0.04	0.40
CGI 11	Reasoning	0.78	0.64	0.13	0.04	0.08	0.74	0.35	0.05	0.04	0.00
CGI 12	Reasoning	0.76	0.62	0.27	-0.03	0.00	0.72	0.36	0.14	-0.01	-0.05
CGI 13	Reasoning	0.71	0.54	0.09	0.05	0.14	0.67	0.30	0.03	0.09	0.01
CGI 14	Reasoning	0.81	0.54	0.36	0.07	-0.05	0.77	0.31	0.19	-0.05	0.03
CGI 15	Speed of processing	0.74	0.32	-0.06	0.17	0.52	0.72	0.15	-0.07	0.36	0.10
CGI 16	Speed of processing	0.69	-0.13	0.25	0.18	0.64	0.68	-0.08	0.14	0.44	0.12
CGI 17	Speed of processing	0.57	0.02	-0.00	-0.14	0.96	0.60	0.02	-0.02	0.68	-0.10
CGI 18	Social Cognition	0.73	0.04	0.74	0.06	0.01	0.70	0.06	0.44	-0.00	0.04
CGI 19	Social Cognition	0.76	0.36	0.63	-0.10	-0.01	0.73	0.24	0.37	-0.02	-0.08
CGI 20	Social Cognition	0.69	-0.16	0.83	0.06	0.10	0.67	-0.05	0.50	0.06	0.05
CGI 21	Social Cognition	0.72	-0.00	0.86	0.02	-0.02	0.69	0.05	0.51	-0.04	0.01
Factor Correlations		1.0									
		.69	1.0								
		.60	.55	1.0							
		.54	.58	.52	1.0						

Note. N = 323. Salient loadings are in bold. Uni means factor loadings in the unidimensional solution; F1 through F4 indicate factor loadings in the oblique factor solution; Gen indicates the general factor loading and GR1 through GR4 indicate group factor loadings in the bifactor solution.

Table 4

Unidimensional, Multidimensional, and Bifactor Analyses of SCoRS

Item	Cognitive Domain	Uni.	Four-Factor Oblique Rotation				Schmid-Leiman Bifactor					
			F1	F2	F3	F4	Gen	GR1	GR2	GR3	GR4	
SCR 1	Memory	0.55	-0.23	0.68	0.23	0.08	0.50	-0.47	0.16	-0.13	-0.06	
SCR 2	Memory	0.65	0.27	0.46	-0.06	0.14	0.57	-0.35	-0.17	0.07	-0.10	
SCR 3	Attention	0.57	0.51	0.06	-0.00	0.12	0.52	-0.05	-0.33	0.01	-0.07	
SCR 4	Memory	0.57	0.06	0.47	0.06	0.14	0.51	-0.34	-0.04	-0.02	-0.10	
SCR 5	Memory	0.59	-0.01	0.27	-0.06	0.66	0.56	-0.17	0.02	0.09	-0.50	
SCR 6	Learning	0.61	0.45	0.28	0.15	-0.17	0.54	-0.22	-0.29	-0.11	0.15	
SCR 7	Problem Solving	0.70	0.27	0.74	-0.09	-0.03	0.62	-0.57	-0.17	0.09	0.04	
SCR 8	W. Mem	0.65	-0.14	0.47	0.59	-0.04	0.61	-0.29	0.11	-0.39	0.05	
SCR 9	Problem Solving	0.58	0.45	0.04	-0.07	0.33	0.54	-0.03	-0.29	0.07	-0.23	
SCR 10	Speed	0.53	0.24	0.22	0.20	-0.00	0.48	-0.16	-0.15	-0.13	0.02	
SCR 11	Attention	0.61	0.58	0.08	0.09	-0.03	0.55	-0.07	-0.38	-0.06	0.05	
SCR 12	Problem Solving	0.54	0.03	-0.03	0.10	0.71	0.54	0.08	-0.01	-0.03	-0.53	
SCR 13	Attention	0.53	0.51	0.05	0.06	0.01	0.48	-0.04	-0.33	-0.04	0.01	
SCR 14	Learning	0.79	0.25	0.42	0.23	0.09	0.71	-0.30	-0.15	-0.14	-0.05	
SCR 15	W. Mem	0.55	-0.04	0.09	0.60	0.10	0.54	-0.01	0.04	-0.40	-0.07	
SCR 16	Speed	0.50	-0.03	0.03	0.35	0.37	0.50	0.03	0.03	-0.21	-0.27	
SCR 17	Prob. Solving	0.63	0.40	-0.09	0.17	0.35	0.61	0.09	-0.25	-0.10	-0.24	
SCR 18	Social Cog.	0.65	0.22	0.11	0.62	-0.13	0.62	-0.04	-0.14	-0.43	0.13	
SCR 19	Social Cog.	0.58	0.21	-0.15	0.57	0.14	0.57	0.17	-0.12	-0.39	-0.08	
SCR 20	Social Cog.	0.49	0.76	-0.13	-0.05	0.00	0.45	0.08	-0.49	0.03	0.02	
Factor Correlations		1.0										
		.54	1.0									
		.51	.44	1.0								
		.49	.35	.44	1.0							

Note. N = 319. Salient loadings are in bold. Uni means factor loadings in the unidimensional solution; F1 through F4 indicate factor loadings in the oblique factor solution; Gen indicates the general factor loading and GR1 through GR4 indicate group factor loadings in the bifactor solution.

Table 5

Item Response Theory Item Parameter Estimates and Standard Errors For the Graded Response Model

Item	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
CGI 1	2.26 (0.22)	-1.93 (0.20)	-0.57 (0.09)	0.60 (0.10)	1.85 (0.21)
CGI 2	2.20 (0.24)	-1.02 (0.12)	-0.37 (0.08)	0.57 (0.11)	1.70 (0.20)
CGI 3	2.02 (0.22)	-1.65 (0.17)	-0.78 (0.11)	0.39 (0.10)	1.86 (0.18)
CGI 4	2.36 (0.25)	-0.77 (0.10)	0.24 (0.09)	1.01 (0.12)	2.24 (0.24)
CGI 5	1.72 (0.19)	-1.87 (0.21)	-0.87 (0.13)	0.43 (0.12)	1.76 (0.22)
CGI 6	2.15 (0.21)	-1.78 (0.19)	-0.81 (0.11)	0.40 (0.10)	1.58 (0.17)
CGI 7	1.72 (0.20)	-0.75 (0.14)	0.40 (0.11)	1.42 (0.17)	2.81 (0.34)
CGI 8	1.64 (0.21)	-0.42 (0.12)	0.74 (0.13)	2.03 (0.25)	3.31 (0.35)
CGI 9	1.45 (0.19)	-0.96 (0.15)	0.27 (0.13)	1.24 (0.18)	2.44 (0.34)
CGI 10	1.56 (0.18)	-0.70 (0.14)	0.37 (0.12)	1.33 (0.18)	2.68 (0.34)
CGI 11	2.22 (0.26)	-1.52 (0.16)	-0.77 (0.10)	0.25 (0.09)	1.36 (0.17)
CGI 12	2.05 (0.23)	-1.26 (0.14)	-0.47 (0.10)	0.61 (0.11)	1.83 (0.19)
CGI 13	1.67 (0.20)	-1.12 (0.14)	-0.15 (0.11)	0.85 (0.13)	1.94 (0.24)
CGI 14	2.41 (0.25)	-1.07 (0.10)	-0.07 (0.09)	0.75 (0.11)	1.58 (0.17)
CGI 15	1.92 (0.21)	-0.84 (0.13)	-0.08 (0.10)	0.81 (0.13)	2.19 (0.24)
CGI 16	1.54 (0.19)	-0.42 (0.13)	0.68 (0.13)	1.57 (0.21)	2.82 (0.37)
CGI 17	1.06 (0.18)	-0.03 (0.16)	1.37 (0.26)	2.72 (0.45)	
CGI 18	1.79 (0.22)	-0.77 (0.13)	-0.08 (0.10)	0.80 (0.13)	2.11 (0.25)
CGI 19	2.01 (0.23)	-1.32 (0.15)	-0.43 (0.10)	0.65 (0.12)	1.71 (0.19)
CGI 20	1.53 (0.18)	-0.97 (0.15)	-0.04 (0.11)	0.89 (0.15)	2.16 (0.28)
CGI 21	1.72 (0.22)	-0.59 (0.13)	0.24 (0.11)	1.24 (0.16)	2.42 (0.28)
SCR 1	0.97 (0.17)	-1.56 (0.29)	1.47 (0.31)		
SCR 2	1.41 (0.20)	0.25 (0.13)	1.88 (0.28)		
SCR 3	1.08 (0.18)	-1.21 (0.24)	0.75 (0.20)	3.42 (0.63)	
SCR 4	1.17 (0.17)	-0.60 (0.17)	1.50 (0.24)		
SCR 5	1.23 (0.19)	-0.57 (0.16)	1.76 (0.26)	3.51 (0.58)	
SCR 6	1.30 (0.19)	-0.87 (0.18)	0.72 (0.17)		
SCR 7	1.61 (0.22)	-1.43 (0.18)	0.63 (0.14)	2.77 (0.34)	
SCR 8	1.34 (0.18)	-0.70 (0.15)	1.64 (0.24)		

Item	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4
SCR 9	1.34 (0.19)	-0.89 (0.17)	0.42 (0.14)	2.77 (0.40)	
SCR 10	0.95 (0.18)	0.03 (0.18)	2.91 (0.51)		
SCR 11	1.25 (0.19)	-1.59 (0.24)	0.09 (0.15)	3.08 (0.53)	
SCR 12	1.17 (0.21)	0.86 (0.19)	2.54 (0.43)		
SCR 13	0.95 (0.16)	-1.99 (0.38)	1.38 (0.26)	3.48 (0.58)	
SCR 14	1.97 (0.25)	-0.72 (0.12)	0.77 (0.13)	3.56 (0.56)	
SCR 15	1.13 (0.17)	-0.51 (0.17)	1.59 (0.28)	3.37 (0.56)	
SCR 16	0.85 (0.17)	-0.40 (0.21)	2.43 (0.47)		
SCR 17	1.48 (0.21)	-0.79 (0.15)	1.04 (0.18)		
SCR 18	1.45 (0.21)	-0.24 (0.13)	1.52 (0.21)	3.01 (0.40)	
SCR 19	1.37 (0.18)	-0.79 (0.16)	0.77 (0.15)	2.66 (0.38)	
SCR 20	1.06 (0.18)	-1.16 (0.24)	1.54 (0.27)	3.79 (0.71)	