Published in final edited form as:

Appl Psychol Meas. 2010 September ; 34(6): 410–423. doi:10.1177/0146621609349802.

A Comparison of Content-Balancing Procedures for Estimating Multiple Clinical Domains in Computerized Adaptive Testing: Relative Precision, Validity, and Detection of Persons With Misfitting Responses

Barth B. Riley, PhD¹, Michael L. Dennis, PhD¹, and Kendon J. Conrad, PhD²

¹Chestnut Health Systems, Bloomington, IL

²University of Illinois at Chicago

Abstract

This simulation study sought to compare four different computerized adaptive testing (CAT) content balancing procedures designed for use in a multidimensional assessment with respect to measurement precision, symptom severity classification; validity of clinical diagnostic recommendations; and sensitivity to atypical responding. The four content balancing procedures were: (1) no content balancing, (2) Screener-based, (3) Mixed (Screener plus content balancing), and (4) Full content balancing. In Full content balancing and in Mixed content balancing following administration of the screener items, item selection was based on: (1) whether the target number of items for the item's subscale was reached, and (2) the item's information function. Mixed and Full content balancing provided the best representation of items from each of the main *IMDS* subscales. These procedures also resulted in higher CAT-to-full-scale correlations for the Trauma and Homicidal/Suicidal Thought subscales, and improved detection of atypical responding.

Keywords

computerized adaptive testing; content balancing; multidimensional assessment; atypical responding; classification; Rasch measurement; diagnostic classification; substance use disorder; Global Appraisal of Individual Needs (GAIN)

Introduction

The ability of computerized adaptive testing (CAT) to significantly reduce assessment time and respondent burden without major loss of measurement precision and validity has led to its growing application in general and in recent years within the area of mental health assessment (Gardner, Kelleher, & Pajer, 2002; Gardner et al., 2004; Gibbons et al., 2008; Handel, Ben-Porath, & Watt, 1999). The application of CAT in clinical assessment has required researchers to address two important issues. First, measures are frequently required across multiple dimensions comprising a second-order factor. CAT item selection procedures

must therefore be tailored to collect sufficient information for measurement estimation on both first- and second-order dimensions. Second, atypical response patterns (e.g., reporting only high severity items like suicidal ideation without reporting low severity items like being depressed) can bias the score downwards and impact interpretation. Thus, if one person endorses three low severity symptoms of depression and another endorses 3 high severity symptoms of trauma or suicidal thoughts – they would nominally get the same severity score (3), but the latter would be considered by most clinicians to be more severely distressed. Previous research suggests that because CAT generally selects items that have an equal probability of endorsement vs. non-endorsement, they are often insensitive in detecting such atypical response patterns (Meijer, Molenaar, & Sijtsma, 1994; Molenaar & Hoijink, 1990, 1996; Reise & Due, 1991).

Assessment on Multiple Clinical Domains

Multidimensional measurement in CAT has traditionally been achieved through sequential administration of unidimensional adaptive tests (Handel et al., 1999). This method, however, has been criticized as inefficient, particularly when subscales are short and obliquely correlated (Wang & Chen, 2004). Recently, researchers have examined the utility of multidimensional Rasch and Item Response Theory (IRT) models in adaptive testing (Briggs & Wilson, 2003; Gardner et al., 2002; Gibbons et al., 2008; Haley, Fragala-Pinkham, & Ni, 2006; Luecht, 1996; Petersen et al., 2006; Segall, 1996, 2001; Tseng, 2001; van der Linden, 1999; Veldkamp, 2002; Veldkamp & van der Linden, 2002; Wang & Chen, 2004; Wang, Chen, & Cheng, 2004). Using computer simulations, Wang and Chen (2004) found that relative to sequentially-administered unidimensional CATs, multidimensional adaptive testing became increasingly more efficient as the number of latent traits, the correlation among latent traits and the number of response categories increased. Similar findings have been reported in other CAT simulation studies (Haley et al., 2006; Segall, 1996).

Despite these promising findings, the increased efficiency of multidimensional adaptive testing may be offset by the following: (1) the increased computer processing time needed in multidimensional trait estimation (thereby leading to unacceptable time delays between questions and in turn causing respondent frustration, particularly when administering a CAT via the Internet); and (2) issues concerning the interpretability of trait estimates derived from multidimensional measurement models. For instance, two persons may obtain different depression measures even though they endorsed the same depression items, because one person endorsed more pain items than the other, and pain being correlated with depression influenced the depression measures differentially (Fayers, 2007). Despite correlations among domains, it is desirable from a clinical perspective to keep domains such as depression and pain distinct in order to facilitate clinical diagnosis and treatment planning (Fayers, 2007). Therefore, alternative approaches to the assessment of individuals on multiple dimensions should be explored.

Identifying Subgroups with Atypical Response Patterns Using CAT

Modern psychometric models predict how persons at a given level on a latent dimension will respond to a set of items that measure that dimension based on a set of item parameters. However, not all people follow this predicted pattern of responding. Atypical response

patterns, or *person misfit* to the measurement model, may have important clinical implications. For instance, work by Conrad and colleagues (Conrad, Bezruczko, Park, Chan, Riley, Diamond et al., 2008; Conrad, Conrad, Dennis and Riley, 2008) has identified subgroups of people with atypical clinical presentations that were consistent in endorsing a set of high severity symptoms, but who did not endorse the expected number of low and moderate severity symptoms leading up to it. Examples of atypical responding include someone presenting with withdrawal related to substance use without having first developed substance dependence (which happens in early opiate or methamphetamine users) or someone presenting with suicide ideation in the absence of depression symptoms. Such atypical patterns alter how we interpret the overall measures and frequently change the expected course of treatment (e.g., using depression treatments for suicide may not make sense if a person does not have depression). They also represent a variation in how the subscales are related to each other and in so doing identify atypical response subgroups. While there has been considerable research concerning the development and sensitivity of various person fit indices in general (Dimitrov & Smith, 2006; Karabatsos, 2003; Li & Olejnik, 1997; Meijer & Sijtsma, 2001) and within the context of CAT (Nering, 1995; van Krimpen-Stoop & Meijer, 1999, 2000, 2001), there has been little research on the effect of CAT item selection procedures on the detection of atypical response patterns. We hypothesize that ensuring adequate measurement of each subdimension will allow CAT to still detect these patterns in spite of using fewer items.

Content Balancing

Content balancing is most often used to ensure that different content areas covered by a test or instrument are adequately represented during an adaptive testing session (Luecht, Champlain, & Nungester, 1998). For example, a math test may consist of arithmetic questions and story problems. Content balancing could be used to ensure that some percentage of the administered items are arithmetic questions while the remaining items administered during the CAT are story problems. In a typical content balancing algorithm, an item is selected according to two criteria: (1) the level of information provided by the item relative to the respondent's ability, and (2) if the content domain of the item has not been satisfactorily covered by the previously administered items.

Content balancing could potentially be applied to the assessment of multiple dimensions in order to ensure adequate coverage and measurement of each subdimension. This is provided that each item loads on a single dimension or on one subdimension and a second-order factor. A potential advantage of this approach is improved detection of persons with random or atypical response patterns *across* the various subdomains, particularly when different subdomains reflect varying levels of symptom severity (e.g., depression vs. suicidal) on the first-order factor.

The purpose of this study was to evaluate the use of content balancing in CAT involving the assessment of multiple dimensions and its effect on measurement and classification precision, accuracy of diagnostic recommendations, and sensitivity to atypical responding. The study addresses the following research questions:

Relative to CAT without content balancing, does content balancing improve

1. the precision of subscale measures and the accuracy of classification by symptom severity (low, moderate, high)?
2. the validity of clinical diagnostic recommendation?
3. the ability to detect atypical response patterns?

Method

Data Source

The data used in this study are from 4,363 people who were interviewed using the *Global Appraisal of Individual Needs*, (Dennis, Titus, White, Unsicker, & Hodgkins, 2003) Internal Mental Distress Scale. The data were collected over 77 studies conducted in three dozen locations around the United States. Over two-thirds of these studies were conducted by independent investigators. All data were collected as part of general clinical practice or specific research studies under their respective voluntary consent procedures, pooled, de-identified, and made available for secondary data analyses by the *GAIN* Coordinating Center at Chestnut Health Systems.

Sample Characteristics

Table 1 presents the demographic and clinical characteristics of the sample. Most of the participants in the study were adolescents (81%), males (69%), and involved in the juvenile/criminal justice system (66%). The sample was 42% Caucasian, with African-Americans representing the next largest ethnic group (23%). Virtually all participants (91%) had some form of psychiatric or substance use disorder or past criminal or violent behavior, with substance use disorders (abuse or dependence) being the most prevalent (81%), followed by crime/violence problems (59%), externalizing disorders (59%) and internalizing disorders (47%).

Instruments

The *Internal Mental Distress Scale (IMDS)* is part of the *Global Appraisal of Individual Needs (GAIN)* and consists of 43 items covering symptoms related to internal psychological distress and four subscales (Depression Symptom Scale, Anxiety Symptom Scale, Traumatic Stress Scale, Homicidal/Suicidal Thoughts Scale). The full *IMDS* also includes 4 questions about somatic symptoms (i.e., mind or body not working) to fill in gaps in the scale and support diagnostic categorization. Several of the *IMDS* items are taken from diagnostic criteria in the *Diagnostic and Statistical Manual, Version Four (DSM-IV*; American Psychiatric Association, 2000) for Mood Disorders (Depression Symptom Scale), Anxiety Disorders (Anxiety Symptom Scale), Acute Stress and Post-Traumatic Stress Disorders, and Disorders of Extreme Stress (Traumatic Stress Scale). In addition to *DSM* criteria, items in the *IMDS* come from the Hopkins Symptom Checklist 90 (HSCL; Derogatis, Lipman, & Covi, 1973; Lipman, Covi, & Shapiro, 1979), Mississippi Scale of PTSD (King, King, Fairbank, Schlenger, & Surface, 1993), and the Chestnut Health System's suicide risk screener. Copies of the items and information concerning the psychometric properties of the *IMDS* are available in a detailed technical report (Conrad, Conrad et al 2008).

The *Internalizing Disorder Screener (IDScr)* consists of a subset of five *IMDS* items, one from each of the subscales and one somatic complaint; The 5-item *IDScr* is highly correlated with the full scale (.89 for adolescents and .9 for adults) (Dennis, Chan & Funk, 2006). In an independent evaluation of the *GAIN Short Screener (GSS)*, which includes the *IDScr* and 3 other screeners (22-item *Mini-International Neuropsychiatric Interview-Modified [MINI-M]*, 24-item *Mental Health Screening Form [MHSF]*, and the 6-item *Co-occurring Disorder Screening Instrument for Mental Disorders [CODSI-MD]*), the *GSS* was found to have the best sensitivity (81.6%) and tied for overall accuracy (72.6%) relative to the *Structured Clinical Interview for DSM-IV (SCID)* (Sacks et al 2007).

Content Balancing Procedures

No Content Balancing—Each item was selected to maximize Fisher's information function (Wainer, 2000). In the simulations presented in this paper, item information is estimated using the item's difficulty parameter and the current estimate of the person's measure on the overall *IMDS* scale. This approach focuses on estimating the overall *IMDS* score without concern for estimating the subscales or atypical response patterns.

Screener Content Balancing—All participants started with the five items from the *Internalizing Disorder Screener (IDScr)*, then additional items were selected that maximized Fisher's information function. Administering items from the screener ensures that at least one item from each *IMDS* subscale is administered.

Mixed Content Balancing—Mixed content balancing also began by administering the five screener items. Afterwards, items were selected based on: (1) whether the target number of items for the item's subscale was reached, and (2) the item's information function. Thus, the item that met the first criterion and had the highest information function was selected.

Full Content Balancing—Full content balancing employed this latter approach throughout the CAT session (not requiring screener items to be included).

Development of Content Balancing Criteria and Stop Rule

To focus the comparison, we wanted to use a common number of items across methods. A series of computer simulations were conducted in order to determine the optimal number of items to administer during the CAT. Specifically, we identified the number of items to administer from each of the *IMDS* subscales when using Mixed and Full content balancing, and identified the total number of items to administer when employing screener and no content balancing. Criteria for determining the target number of items were: (1) mean root mean squared error (RMSE) across content-balancing methods < 0.60, and (2) mean Cohen's *d* for measurement bias, (mean CAT measure minus mean full scale measure divided by full scale standard error) falling between $-.20$ and $.20$. These simulations revealed that 18 items (5 from Depression, 5 from Anxiety, 5 from Trauma, and 3 from Homicidal/Suicidal Thoughts in Mixed and Full procedures) would meet the above criteria with minimal loss in measurement efficiency. Since the *IMDS Screener* includes one item from the *Somatic Symptoms Index*, 19 items were administered during the Screener and Mixed content balancing procedures. All CAT simulations were fixed length; that is, the stop

rule for the CAT was based on the total number of administered items (Screener and no content balancing) or administering the targeted number of items from each *IMDS* subscale (Mixed and Full content balancing).

CAT Simulation and Missing Data Procedures

We developed a computer program to perform real-data CAT simulations using each of the content balancing procedures described above. This program used item calibrations that were previously estimated in *Winsteps* (Linacre, 2006) and it calculated person measures using algorithms of the Rasch dichotomous model described by Wright and Douglas (1996). Only data collected at intake were used during the simulations.

When no items for a given subscale were administered, two methods of handling the missing observations were examined: (1) simply treat the cases with no observations as missing data, or (2) replace the missing value with the overall *IMDS* measure based on the administered items. We found little difference in the simulation results between these two methods. Since the latter method (missing value replacement) would most likely be used in a clinical setting, this procedure was employed when reporting the simulation results.

CAT Evaluation Criteria

Measurement Precision and Classification Agreement—We assessed the results of the CAT using several established computerized adaptive testing criteria (Wainer, 2000). Precision was measured in terms of bias, the mean standard error of measurement (MSE), and the root mean square error (RMSE). Bias is calculated by subtracting the mean of the CAT-derived measure from the mean of the measure based on all of the items, and can indicate whether the CAT is overestimating or underestimating trait levels relative to the full instrument. Ideally, bias should be equal to zero. Measurement error decreases as the targeting of items to the person's trait level increases (Bond and Fox, 2007). Mean standard error (MSE) therefore reflects the overall targeting of the items across all trait estimates, with lower values indicating greater measurement precision. RMSE is calculated by computing the square root of the average squared differences between each respondent's CAT-estimated measure and their measure based on the full instrument. The RMSE can be thought of as the standard deviation of the differences between the CAT and full-scale measurements, with lower RMSE values indicating greater consistency between measures derived from the CAT compared with measures derived from the full measure. Kappa coefficients were used to assess the agreement between triage of symptom severity based on the CAT measures and triage based on the full *IMDS*.

Validity—As a preliminary analysis of the validity of the CAT-derived measures, receiver operator curve (ROC) analyses were performed to predict meeting of *DSM* diagnostic criteria of major depression, generalized anxiety and traumatic stress disorder based on CAT-derived *IMDS* subscale measures of depression, anxiety and trauma, respectively. As the *IMDS* items are based directly on *DSM* criteria, respondents were classified as meeting diagnostic criteria based on their responses to specific *IMDS* items.

Results

Does content balancing improve the precision of subscales as continuous measures and triage groups (low, moderate, high)?

Table 2 presents mean precision (bias, MSE and RMSE) for the four content balancing procedures (None, Screener, Mixed and Full). The amount of bias ranged from -0.808 to $+0.076$. The Screener procedure had the highest average bias (-0.222) while the Mixed and Full content balancing procedures did the best on average. No content balancing and the Screener procedures were biased by more than $-.2$ on trauma and homicidal/suicidal thought and Mixed balancing was biased over $.2$ on trauma. Only the Full content balancing resulted in no absolute bias over $.2$. The particular large bias of the screener method on the homicidal/suicidal thought subscale (bias $d = -0.808$) suggests that when the screener is used for content balancing, the CAT fails to select more severe items that the respondent would have endorsed. The Mean SE did not change markedly across the four content balancing procedures. Relative to the full-instrument SEs, however, markedly larger MSEs (i.e., 20% larger than mean full-instrument MSE) were observed for the following content balancing procedures and *IMDS* subscales: No content balancing: *Depression* (MSE=1.40); Screener: *Depression* (MSE=1.38); Mixed content balancing: *IMDS* scale (MSE=0.85); and Full content balancing: *IMDS* scale (MSE=0.88). The Mixed and Full content balancing procedures produced much smaller RMSE values, with all $.62$ or lower. The no content balancing condition had problems with *Homicidal/Suicidal Thoughts* and the Screener condition had problems with that and *Depression*.

Table 3 presents the CAT to full scale correlations by content balancing condition for the and subscales. All procedures produced high correlations with the full *IMDS* scale ($r=.971$ to $.982$), *Depression* subscale ($r=.937$ to $.957$), *Anxiety* subscale ($r=.947$ to $.962$), and *Trauma* subscale ($r=.968$ to $.974$). Content balancing had the largest impact on the *Homicidal/Suicidal Thoughts* scale; when no content balancing was employed, the CAT to full-scale correlation was $.599$, whereas the correlation for the Screener content balancing condition was $.828$ and was even better for Mixed ($r=.964$) and Full ($r=.945$) content balancing. The average CAT to full-scale correlation was highest for Mixed content balancing ($r=.965$) and lowest for No Content Balancing ($r=.894$).

We examined the ability of the *IMDS* CAT to place individuals into one of three symptom severity groups (low, moderate, and high). Classification of individuals into triage groups was based on the Rasch measure for the *IMDS* scale and the four *IMDS* subscales using the cut points shown in Table 4. The Rasch cut points in turn were set on the basis of pre-existing raw score cut points. The *GAIN IMDS* raw score range is from 0 to 43, with scores categorized as low (0 to 8), moderate (9 to 23) or high (24 to 43). The raw score and CAT score versions of *IMDS* were converted to a common Rasch metric to allow for direct comparisons and create a continuous measure with mean of 0 and a standard deviation of 1 logit. The cutpoints on the Rasch metric were set based on where the majority of individuals fell for each of the three categorization. Thus, the low-to-moderate cutpoint corresponded to 90% of the individuals below it who had a raw *IMDS* scores in the low range (0 to 8). Conversely, the moderate-to-high cutpoint was set so that 90% of the individuals above it

had a raw *IMDS* scores in the high range (24 to 43). The middle group had a mix of all three raw score groups with the majority falling in the moderate range (9 to 23). This procedure has already been described at length elsewhere (Riley, Conrad, Bezruczko, & Dennis, 2007).

To the right of these cut points in Table 4 are the Kappa coefficients indicating the level of agreement between the full-measure triage classifications and each of the four CAT content balancing procedures. Kappa are all in the excellent range for the overall *IMDS* ($K = .841$ to $.871$), *Depression* subscale ($K = .749$ to $.911$), *Anxiety* subscale ($K = .759$ to $.811$), and *Trauma* ($K = .836$ to $.850$). For the *Homicidal/Suicidal Thoughts* subscale, the kappas were excellent for Mixed ($K = .917$) and Full ($K = .902$) content balancing but unacceptably low for the no balancing ($K = .312$) and Screener ($K = .067$) procedures. When averaging the results across all *IMDS* measures, Mixed and Full content balancing provided substantially better classification ($K = .838$ and $.824$ respectively) compared to No and Screener Content Balancing ($K = .745$ and $.692$, respectively).

Does content balancing improve the validity of clinical diagnostic recommendations?

A series of receiver operator curve (ROC) analyses were performed using SPSS version 16.0 (Chicago, Illinois) to predict *DSM-IV* diagnostic recommendations (based on *DSM-IV* algorithms) for major depression, generalized anxiety and traumatic stress, using CAT-based *IMDS* measures of depression, anxiety and trauma as predictors, respectively. The *DSM* diagnostic algorithms impose rules requiring presence of specific symptoms or combination of symptoms for a diagnosis to be made. As shown in Table 5, all area under curve (AUC) for predicting a diagnostic recommendation using measures based on the different content balancing procedures were in the .91-.99 range and were generally similar across procedures. Sensitivity ranged from .88 to .97, and specificity ranged from .86 to .96. Using the full Rasch instrument to predict *DSM* diagnostic recommendations, AUCs ranged from .96 to .99, sensitivity ranged from .91 to 1.00, and specificity ranged from .89 to 1.00. Thus, CAT-derived measures were nearly as predictive of diagnostic criteria as measures obtained from the full *IMDS*. Both CAT and the full-instrument measures provided equal and in some cases better prediction of diagnostic impressions than symptom counts.

Does content balancing improve the ability to detect atypical response patterns?

Next, we examined the ability of content balancing to improve identification of persons with atypical response patterns at the *IMDS* scale level. Table 2 presents correlations between CAT-derived and full-instrument fit statistics by content balancing condition using the following fit indices: (1) mean square outfit (Wright & Stone, 1979); (2) mean square infit (Wright & Stone, 1979); (3) standardized outfit (Smith, Schumacker, & Bush, 1998); and (4) E_i (Kane & Brennan, 1980). Note that correlations between CAT and full-instrument indices were examined rather than ROC analyses due to the absence of a “gold standard” statistic and the fact that not all fit statistics have established cut-off values. Outfit, t , and E_i were selected due to their sensitivity to outlying aberrant response patterns (i.e., persons with low trait levels endorsing high difficulty/severity items (Dimitrov & Smith, 2006; Karabatsos, 2003) consistent with the atypical suicide pattern. Infit was also examined since inlying aberrant responses can affect measurement precision (Wright, 1995). Those with correlations under .7 are bolded. As seen in Table 2, correlations between CAT and full-

instrument outfit, standardized outfit (t) and Ei were highest when Mixed content balancing was employed. Conversely, correlations for infit were highest for the Screener and no content balancing procedures.

Discussion

The results of the present study indicate that content balancing was generally effective in providing precise assessment of multiple clinical domains, accurate prediction of diagnostic criteria and improved detection of atypical response patterns. Specifically, content balancing (especially Mixed and Full content balancing) were successful in reducing measurement bias, improving average measurement precision (as measured by RMSE) and classification accuracy across most *IMDS* subscales. The latter finding was particularly true for the *Homicidal/Suicidal Thoughts* subscale. Conversely, content balancing did not improve the measurement precision and classification accuracy for the other *IMDS* subscales relative to the no content balancing condition. However, these results indicate that the benefits of content balancing associated with the *Homicidal/Suicidal Thoughts* subscale outweighs the relatively small reductions in precision and classification accuracy observed on the other *IMDS* subscales. When the *Homicidal/Suicidal Thoughts* subscale is not taken into consideration, the improvement in mean CAT-to-full-scale correlations resulting from the No Content Balancing relative to the Mixed and Full content balancing procedures is .001 and .004, respectively. When the *Hostility/Suicidality* subscale is considered, the difference between No Content Balancing and Mixed and Full content balancing are 0.71 and 0.66, respectively.

To explain these findings, it is important to note that, unlike the other *IMDS* subscales that have items well distributed across the *IMDS* continuum, the *Homicidal/Suicidal Thoughts* subscale reflects a specific (i.e., high) level of internal mental distress. Thus, when the no content balancing condition is employed, *Homicidal/Suicidal Thoughts* items may not be administered during the CAT if the overall *IMDS* measure is in the low to average range (note that item selection across all content balancing methods was based on the overall *IMDS* measure). This, coupled with the fact that the *Homicidal/Suicidal Thoughts* subscale consists of relatively few (n=5) items, would seem to explain why content balancing provides the greatest benefit for this measure.

The results of the present study also suggest that CAT-derived measures of depression, anxiety and traumatic stress were associated with suggestive *DSM* diagnoses of major depression, generalized anxiety, and traumatic stress disorder, respectively. There was little difference in predictive efficacy between content balancing procedures, with CAT providing near to equally effective prediction of diagnostic criteria relative to the full instrument. It should be noted, however, that *DSM* diagnoses were not made by an independent clinician/assessment protocol but were based on responses to the *IMDS* items used to estimate the CAT-derived measures. The present findings may therefore overestimate the predictive efficacy of the CAT when compared to standard clinical practice. Gardner et al., (2004) for instance reported an area under the curve (AUC) of .88 when predicting independent diagnoses of major depression based on adaptively-administered *Beck*

Depression Inventory measures. In contrast, ROC analyses in the present study revealed AUCs of .97-.98 for suggestive diagnosis of major depression.

Despite considerable research concerning the development of person fit indices there has been little empirical work concerning the effect of CAT item selection on the sensitivity of person fit statistics. Relative to the full instrument, the present study suggests that content balancing can improve the sensitivity of CAT-based person fit measures designed to detect atypical, *outlying* aberrant response patterns (e.g., persons with low measures on the latent trait endorsing high severity items). Specifically, CAT to full-instrument correlations of outfit, standardized outfit and E_i improved when using Mixed and Full content balancing compared to no content balancing. Conversely, Mixed and Full content balancing reduced the sensitivity of fit indices designed to detect random, *inlying* aberrant response patterns (i.e., as measured by infit). This is likely due to the fact that content balancing effectively requires administration of items across the measurement spectrum, whereas detection of random inlying patterns requires administration of items near the respondent's *IMDS* measure. As has been suggested, however, detection of outlying aberrant response patterns may have important clinical implications with respect to treatment placement.

Strengths and Limitations

While this study had many strengths (large and diverse sample, testing several content balancing procedures, using several types of construct validity), it does have some important limitations that need to be acknowledged. These include: (1) a limited item bank size, (2) the use of simulation procedures which may not accurately reflect how persons respond to a working CAT application, (3) *DSM* diagnoses not based on an independent clinical evaluation, and (4) no control for the small amounts of differential item functioning (DIF) by gender, age and race detected in the *IMDS* instrument (Conrad, Conrad et al., 2008) although DIF was not the focus of the present study. It should also be noted that because a fixed-length stop rule was employed, we could not directly assess the relative efficiency of the four content balancing procedures.

The present study generally supports the utility of CAT content balancing methods to improve overall measurement precision on multiple clinical domains, accuracy of triage of persons by symptom severity, and the detection of persons with atypical presentation of symptoms. Preliminary evidence suggests that CAT-based measures can be used for accurate determination of diagnostic status. However, future research to independently validate these procedures against actual clinical diagnosis and treatment placement is warranted.

Acknowledgments

This development of this paper was supported by the Center for Substance Abuse Treatment (CSAT), Substance Abuse and Mental Health Services Administration (SAMHSA) contract 270-07-0191 using data provided by the following grants and contracts from CSAT (TI11317, TI11321, TI11323, TI11324, TI11422, TI11423, TI11424, TI11432, TI11433, TI11871, TI11874, TI11888, TI11892, TI11894, TI113190, TI113305, TI113308, TI113309, TI113313, TI113322, TI113323, TI113340, TI113344, TI113345, TI113354, TI113356, TI113601 TI114090, TI114103, TI114188, TI114189, TI114196, TI114214, TI114252, TI114254, TI114261, TI114267, TI114271, TI114272, TI114283, TI114311, TI114315, TI114355, TI114376, TI115413, TI115433, TI115447, TI115461, TI115467, TI115475, TI115478, TI115479, TI115481, TI115483, TI115485, TI115486, TI115511, TI115514, TI115545, TI115562, TI115670, TI115671, TI115672, TI115674, TI115678, TI115682, TI115686; Contract 207-98-7047, Contract 277-00-6500, Contract 270-2003-00006),

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), the Illinois Office of Alcoholism and Substance Abuse (PI 00567), and the Robert Wood Johnson Foundation's Reclaiming Futures (45054, 45059, 45060, 45053, 047266). The authors would like to thank Dr. Karen Conrad who provided input into the manuscript and contributed to psychometric analyses of the *IMDS*; Rodney Funk who helped with data management; Joan Unsicker who helped to edit the manuscript; and the study staff and participants for their time and effort. The opinions are those of the author and do not reflect official positions of the contributing project directors or government.

References

- American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-IV-TR). 4th. Washington D.C: American Psychiatric Association; 2000.
- Bond, TG., Fox, CM. Applying the Rasch model: Fundamental measurement in the human sciences. 2nd. Mahwah, N.J.: Lawrence Erlbaum; 2007.
- Briggs DC, Wilson M. An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*. 2003; 4(1):87–100. [PubMed: 12700433]
- Conrad KJ, Bezruczko N, Park H, Chan YF, Riley BB, Diamond G, et al. Atypical suicide risk using Rasch person fit statistics to improve screening and understanding. 2008 Manuscript in preparation.
- Conrad, KJ., Conrad, KM., Dennis, ML., Riley, BB. GAIN working papers: Rasch analysis of the Internal Mental Distress Scale. 2008. Retrieved 10/22/2008 from http://www.chestnut.org/li/gain/psychometric_reports/gain_working_papers_rasch_analysis_of_the_internal_mental_distress_scale.pdf
- Dennis ML, Chan YF, Funk R. Development and validation of the GAIN Short Screener (GSS) for internalizing, externalizing, and substance use disorders and crime/violence problems among adolescents and adults. *American Journal on Addictions*. 2006; 15(80–91)
- Dennis, ML., Titus, JC., White, MK., Unsicker, J., Hodgkins, D. Global Appraisal of Individual Needs: Administration Guide for the GAIN and Related Measures. Bloomington, IL: Chestnut Health Systems; 2003.
- Derogatis LR, Lipman RS, Covi L. SCL-90: An outpatient psychiatric rating scale—Preliminary report. *Psychopharmacology Bulletin*. 1973; 9:13–28. [PubMed: 4682398]
- Dimitrov DM, Smith RM. Adjusted Rasch person fit statistics. *Journal of Applied Measurement*. 2006; 7(2):170–183. [PubMed: 16632900]
- Fayers PM. Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. *Quality of Life Research*. 2007; 16(Suppl 1):187–194. [PubMed: 17417722]
- Gardner W, Kelleher KJ, Pajer KA. Multidimensional adaptive testing for mental health problems in primary care. *Med Care*. 2002; 40(9):812–823. [PubMed: 12218771]
- Gardner W, Shear K, Kelleher KJ, Pajer KA, Mammen O, Buysse D, et al. Computerized adaptive measurement of depression: a simulation study. *BMC Psychiatry*. 2004; 4(1):13. [PubMed: 15132755]
- Gibbons RD, Weiss DJ, Kupfer DJ, Frank E, Fagioli A, Grochocinski VJ, et al. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*. 2008; 59(4):361–368. [PubMed: 18378832]
- Haley SM, Fragala-Pinkham M, Ni P. Sensitivity of a computer adaptive assessment for measuring functional mobility changes in children enrolled in a community fitness programme. *Clinical Rehabilitation*. 2006; 20(7):616–622. [PubMed: 16894805]
- Handel RW, Ben-Porath YS, Watt M. Computerized adaptive assessment with the MMPI-2 in a clinical setting. *Psychological Assessment*. 1999; 11(3):369–380.
- Kane MT, Brennan RL. Agreement coefficients as indices of dependability for domain referenced tests. *Applied Psychological Measurement*. 1980; 4:105–126.
- Karabatsos G. Comparing the aberrant response detection performance of thirty-six person fit statistics. *Applied Measurement in Education*. 2003; 16(4):277–298.
- King DW, King LA, Fairbank JA, Schlenger WE, Surface CR. Enhancing the precision of the Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: An application of item response theory. *Psychological Assessment*. 1993; 5:457–471.

- Li MF, Olejnik S. The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*. 1997; 21:215–231.
- Linacre, JM. Winsteps Rasch measurement computer program (Version X.X) [Software]. 2006. Available from <http://www.winsteps.com/>
- Lipman RS, Covi L, Shapiro AK. The Hopkins Symptom Checklist (HSCL): factors derived from the HSCL-90. *Journal of Affective Disorders*. 1979; 1:9–24. [PubMed: 162184]
- Luecht RM. Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*. 1996; 20(4):389–404.
- Luecht RM, Champlain AD, Nungester RJ. Maintaining content validity in computerized adaptive testing. *Advances in Health Sciences Education*. 1998; 3(1):29–41. [PubMed: 12386393]
- Meijer RR, Molenaar IW, Sijtsma K. Item, test and group characteristics and their influence on nonparametric appropriateness measurement. *Applied Psychological Measurement*. 1994; 18:111–120.
- Meijer RR, Sijtsma K. Methodology review: Evaluating person fit. *Applied Psychological Measurement*. 2001; 25(2):107–135.
- Molenaar IW, Hoijink H. The many null distributions of person fit indices. *Psychometrika*. 1990; 55(1):75–106.
- Molenaar IW, Hoijink H. Person fit and the Rasch model, with an application of knowledge of logical quantors. *Applied Measurement in Education*. 1996; 9:27–45.
- Nering ML. The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement*. 1995; 19:121–129.
- Petersen MA, Groenvold M, Aaronson N, Fayers P, Sprangers M, Bjorner JB. Multidimensional computerized adaptive testing of the EORTC QLQ-C30: Basic developments and evaluations. *Quality of Life Research*. 2006; 15(3):315–329. [PubMed: 16547770]
- Reise SP, Due AM. The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*. 1991; 15:217–226.
- Riley BB, Conrad KJ, Bezruczko N, Dennis ML. Relative precision, efficiency and construct validity of different starting and stopping rules for a computerized adaptive test: The GAIN Substance Problem Scale. *Journal of Applied Measurement*. 2007; 8(1)
- Sacks, S., Melnick, G., Coen, C., Banks, S., Friedmann, PD., Grella, C., Knight, K., Zlotnick, C. CJDATS Co-occurring Disorders Screening Instrument (CODSI) for Mental Disorders (MD): A validation study. In: Simpson, DD., Knight, K., editors. *Criminal Justice & Behavior, Special Issue: Offender Needs & Functioning Assessments*. Vol. 34. 2007. p. 1198–1215.
- Segall DO. Multidimensional adaptive testing. *Psychometrika*. 1996; 61(2):331–354.
- Segall DO. General ability measurement: An application of multidimensional item response theory. *Psychometrika*. 2001; 66(1):79–97.
- Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*. 1998; 2(1):66–78. [PubMed: 9661732]
- Tseng FL. Multidimensional adaptive testing using the weighted likelihood estimation. Dissertation Abstracts International Section A: Humanities & Social Sciences. 2001; 61(12-A):4746.
- van der Linden WJ. Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion. *Journal of Educational and Behavioral Statistics*. 1999; 24(4):398–412.
- van Krimpen-Stoop EM, Meijer RR. The Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests. *Applied Psychological Measurement*. 1999; 23(4):327–345.
- van Krimpen-Stoop, EM., Meijer, RR. Detecting person-misfit in adaptive testing using statistical process control techniques. In: van der Linden, WJ., Glas, CAW., editors. *Computer adaptive testing: Theory and practice*. Boston: Kluwer Academic; 2000.
- van Krimpen-Stoop EM, Meijer RR. CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*. 2001; 26(2):199–218.
- Veldkamp BP. Multidimensional constrained test assembly. *Applied Psychological Measurement*. 2002; 26(2):133–146.
- Veldkamp BP, van der Linden WJ. Multidimensional adaptive testing with constraints on test content. *Psychometrika*. 2002; 67(4):575–588.

- Wainer H. Computerized adaptive testing: A primer (2nd ed.). Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers. 2000; 335
- Wang WC, Chen PH. Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*. 2004; 28(5):295–316.
- Wang WC, Chen PH, Cheng YY. Improving Measurement Precision of Test Batteries Using Multidimensional Item Response Models. *Psychological Methods*. 2004; 9(1):116–136. [PubMed: 15053722]
- Wright BD. Which standard error? *Rasch Measurement Transactions*. 1995; 9(2):436.
- Wright BD, Douglas GA. Estimating measures with known item difficulties. *Rasch Measurement Transactions*. 1996; 10(2):499.
- Wright, BD., Stone, MH. Best test design. Chicago: University of Chicago, MESA Press; 1979.

Table 1

Sample Characteristics Based on Full GAIN Assessment

Characteristic	(n=4,363)
Age - Mean (SD)	19 (7.9)
Age Group	
Adolescents	81%
Adults	19%
Male - %	69%
Race-Ethnicity: African American - %	23%
Caucasian - %	42%
Hispanic - %	13%
Mixed/Other - %	22%
Current Involvement in Juvenile/Criminal Justice System	66%
Any Past Year Disorder ^a - %	91%
Any Past Year Internalizing Disorder - %	47%
Depression - %	41%
Anxiety - %	16%
Traumatic Distress ^b - %	27%
Suicidal ^c - %	15%
Any Past Year Externalizing Disorder - %	59%
Attention Deficit/Hyperactivity Disorder - %	43%
Conduct Disorder - %	51%
Any Past Year Substance Use Disorder - %	81%
Dependence - %	54%
Abuse - %	27%
Any Past Year Crime/Violence Problems - %	59%
High Physical Conflict ^d - %	18%
Mod/High General Crime ^e - %	55%

^aIncludes any of the diagnoses listed under the four domains below

^bIncludes Post-Traumatic Stress Disorder, Acute Trauma Disorder and disorders associated with Child Maltreatment/Interpersonal Violence that may be complex and both in the past and ongoing

^cAny suicidal thought, plus means or attempts

^dPhysical violence towards others based on the GAIN's victimization subscale of the Conflict Tactic Scale

^eAny illegal activity across the FBI's Uniform Crime Report categories beyond simple possession/use

Table 2

IMDS CAT Efficiency and Precision and Correlations between Full Instrument and CAT Person Fit Indices by Content Balancing Method

Criteria/Subscales	Content Balancing Methods			
	None	Screeners	Mixed	Full
<u>Bias (average)^a</u>	-0.126	-0.222	-0.071	-0.068
IMDS	0.023	0.041	-0.004	-0.019
Depression	-0.050	0.076	0.002	0.004
Anxiety	-0.079	-0.129	-0.077	-0.087
Trauma	-0.289	-0.292	-0.221	-0.165
Homicide/Suicide	-0.235	-0.808	-0.054	-0.073
<u>Mean SE (average)^b</u>	1.40	1.47	1.40	1.41
IMDS	0.81	0.82	0.85	0.88
Depression	1.40	1.38	1.33	1.33
Anxiety	1.37	1.44	1.40	1.39
Trauma	1.62	1.60	1.59	1.60
Homicide/Suicide	1.79	2.09	1.85	1.85
<u>RMSE (average)^c</u>	0.76	0.72	0.53	0.56
IMDS	0.40	0.40	0.44	0.50
Depression	0.62	0.73	0.61	0.61
Anxiety	0.55	0.65	0.59	0.58
Trauma	0.68	0.62	0.60	0.62
Homicide/Suicide	1.55	1.19	0.41	0.50
<u>Person Fit Indices^d (average)</u>	.533	.684	.751	.728
Outfit	.452	.631	.738	.737
Infit	.880	.904	.847	.813
t (Std. outfit)	.306	.532	.728	.723
E _i	495	670	691	640

^aNote Absolute bias values > .2 in bold

^bMean SE values > (full-instrument SE * 1.2) highlighted in bold

^cRMSE values > .7 highlighted in bold

^dCorrelations with full-instrument person fit indices. Correlations < .7 highlighted in bold

Table 3

Correlations between *IMDS* Measures based on the Full Instrument and Measures Derived from each of the CAT Content Balancing Procedure

IMDS Measure	Content Balancing Procedure			
	None	Screeners	Mixed	Full
IMDS	0.982	0.982	0.978	0.971
Depression	0.957	0.937	0.956	0.956
Homicidal/Suicidal	0.599	0.828	0.964	0.945
Anxiety	0.962	0.947	0.956	0.957
Trauma	0.968	0.974	0.972	0.969
Average r	0.894	0.934	0.965	0.960

Note: Subscale measures in which no items from a subscale were administered during the CAT session were substituted with the overall IMDS measure. Correlations < .9 highlighted in bold.

Table 4
Triage Cut Points and Kappa Coefficients Comparing Triage Classification Derived from Full and CAT *IMDS* Measures

Measure	Cut Points ^a			Kappa Coefficients ^b			
	Low-Med.	Med.-Hi	None	Screeners	Mixed	Full	
IMDS	-1.72	0.28	.867	.871	.863	.841	
Depression	-1.97	0.21	.909	.911	.753	.749	
Homicidal/Suicidal	-2.78	1.30	.312	.067	.917	.902	
Anxiety	-2.35	0.40	.803	.759	.811	.790	
Trauma	-3.74	-0.87	.836	.850	.847	.837	
Average			.745	.692	.838	.824	

^aNotes Cut points on Rasch measure used to triage into low/moderate/high symptom severity.

^bKappa coefficients < .7 highlighted in bold

ROC Area Under Curve (AUC), Sensitivity (Sens.) and Specificity (Spec.) Predicting DSM Diagnoses by *IMDS* Subscale by Content Balancing Method and Full-Instrument Measures

Table 5

Variable	DSM-IV Diagnostic Recommendation									
	Major Depression ^a					Generalized Anxiety ^b				
	AUC	Sens.	Spec.	AUC	Spec.	AUC	Sens.	Spec.	AUC	Spec.
<u>CAT Content Balancing</u>										
None	.98	.95	.93	.96	.92	.90	.99	.96	.93	.93
Screeners	.98	.94	.94	.95	.91	.86	.99	.93	.93	.96
Mixed	.97	.93	.91	.95	.88	.90	.99	.97	.95	.95
Full	.97	.91	.91	.96	.92	.91	.99	.97	.94	.94
<u>Full IMDS</u>										
Rasch measure	.99	.97	.94	.96	.91	.89	1.00	1.00	1.00	1.00
Raw score	.98	.89	.92	.94	.92	.82	1.00	1.00	1.00	1.00

^aNote predicted by IMDS Depression Symptom Scale

^bpredicted by IMDS Anxiety Symptom Scale

^cpredicted by Traumatic Stress Scale