Graduate
Management
Admission
Council®

Patient-Reported Outcomes Measurement and Computerized Adaptive Testing: An Application of Post-Hoc Simulation to a Diagnostic Screening Instrument

Jason C. Immekus
California State University, Fresno
Robert D. Gibbons
Center for Health Statistics, University of Illinois at Chicago
A. John Rush
University of Texas Southwestern Medical Center at Dallas

Presented at the New CAT Models Paper Session, June 8, 2007

Abstract

This study was designed to expand the use of computerized adaptive testing (CAT) to multidimensional patient-reported outcomes measures. A post-hoc simulation CAT administration of the Psychiatric Diagnostic Screening Questionnaire (PDSQ; Zimmerman & Mattia, 2001) was conducted by conceptualizing the instrument's factor structure in terms of the bifactor item response theory (IRT) model (Gibbons & Hedeker, 1992), using the data of 3,791 individuals with major depressive disorder. Comparison of IRT models supported the scale's bifactor structure. Based on the bifactor model, post-hoc simulation administration of the PDSQ was first conducted on the primary dimension and subsequently on each secondary dimension. Results indicated that full and CAT θ estimates on the primary dimension were moderately to highly correlated; however, CAT θs were underestimated for respondents with sparse item endorsements. Comparable full and CAT θ estimates were observed on the secondary dimensions. Implications and suggestions for future research are discussed.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Immekus, J. C., Gibbons, R.D., & Rush, J. A. (2007). Patient-reported outcomes measurement and computerized adaptive testing: An application of post-hoc simulation to a diagnostic screening instrument. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Jason C. Immekus, California State University,
5005 N. Maple Ave., M/S ED303, Fresno, CA 93740.
Email: jimmekus@csufresno.edu.

Patient-Reported Outcomes Measurement and Computerized Adaptive Testing: An Application of Post-Hoc Simulation to a Diagnostic Screening Instrument

Theoretical Framework

A central component to an efficient, cost-effective, and evidence-based health care system is the availability of psychometrically rigorous patient reported outcomes (PRO) measurements. This form of assessment refers to the use of patients' evaluation of their own physical and emotional well-being, generally in response to medical care that they are receiving for treatment purposes. A potential contribution of PRO instruments in this setting includes evaluating (1) patient symptoms, (2) a patient's perspective on treatment, and (3) the efficacy of treatment. Despite the host of measures available to measure various health outcomes (e.g., post-traumatic growth, depression, or mood), an important line of research in this area is investigating the contributions that multidimensional item response theory (IRT) models [e.g., the bifactor model (Gibbons & Hedeker, 1992)] and computerized adaptive testing (CAT) have to offer PRO measurement.

Item response theory (IRT) marks a notable advancement in the modeling of scale data. However, an inherent problem in the use of unidimensional IRT models in mental health measurement is the apparent multidimensional nature of employed scales. In part, however, this multidimensionality is produced by the sampling of items from multiple domains of an overall psychological construct (e.g., depression). For example, in the measurement of quality of life, items are selected from satisfaction domains such as satisfaction with family, income, and neighborhood. It is quite natural for such data to appear to be multidimensional, when in fact, they measure a unidimensional construct (i.e., quality of life); however, the items within domains are more highly correlated than items between domains. This circumstance leads to violation of the conditional independence assumption of a unidimensional model and results in dimensionality equal to the number of domains from which the items were sampled. To apply traditional unidimensional IRT models (e.g., one- or two-parameter models) to such data requires fitting as many models to the data as there are factors underlying the data.

As an alternative, however, a plausible s-factor solution for many mental health measurement scales is one that exhibits a general factor and $s-1$ group, or method, factors. The bifactor solution constrains each item j to have a non-zero loading α_{j1} on the primary dimension and a second loading ($\alpha_{jk}, k = 2,\ldots, s$) on not more than one of the $s-1$ group factors. For four items, the bifactor pattern matrix might be

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & 0 & \alpha_{33} \\ \alpha_{41} & 0 & \alpha_{43} \end{bmatrix} \qquad (1)$$

Holzinger and Swineford (1937) termed this as a “bi-factor” solution. Although the model was originally conceived in terms of continually distributed test scores, it is easy to conceive of situations where the bifactor pattern might also arise at the item level (Muthén, 1989). It is plausible for paragraph comprehension items in achievement tests, for example, where the primary dimension describes the targeted process skill and additional factors describe content area knowledge within paragraphs. Similarly, in the context of mental health measurement, symptom items are often selected from measurement domains and can be related to the primary dimension of interest (e.g., depression or mental instability) and one sub-domain (e.g., anxiety). In these contexts, items would be conditionally independent between paragraphs or domains, but conditionally dependent within paragraphs or domains.

Gibbons and Hedeker (1992) derived an item response model for binary response data exhibiting the bifactor structure and developed a practical parameter estimation method. As they demonstrated, the bifactor restriction leads to a major simplification of likelihood equations that (1) permits analysis of models with large numbers of group factors (e.g., domains), (2) permits conditional dependence among identified subsets of items, and (3) in many cases provides a more parsimonious factor solution than an unrestricted full-information item factor analysis (e.g., Bock & Aitkin, 1981).

Traditional unidimensional IRT models (one-, two, or three-parameter models) have generally served as the basis for computerized adaptive testing (CAT). The basis of CAT is the use of a large, calibrated item bank, with each item selected that provides the most information regarding a respondent’s trait level based on each item response. Therefore, administered items are those that are most appropriate for each respondent (Weiss, 1985). This is contrary to traditional pencil-paper format tests, which require all respondents to complete every scale item. It is typically found that adaptive tests result in a 50% average reduction in number of items administered, and some reductions in the range of 80% to 90% have been reported, with no decrease in measurement quality (Brown & Weiss, 1977; Weiss & Gibbons, 2007). In addition, as has been indicated, adaptive tests allow control over measurement precision. Thus, adaptive tests result in measurements that are both efficient and effective.

Within PRO measurement, both IRT—particularly the bifactor model—and CAT have direct implications for health outcomes measurement. The bifactor model provides a method for modeling the multidimensionality generally displayed by PRO measurements (Chen, West, & Sousa, 2006; Gibbons, Bock, Hedeker, et al., 2007; Reise, Morizot, & Hays, in press). Specific advantages of CAT include, among others: (1) shorter, quicker tests; (2) improved test security, since all items are not administered to every respondent; (3) immediate scoring; and (4) sampling items from an item bank

The Psychiatric Diagnostic Screening Questionnaire (PDSQ; Zimmerman & Mattia, 2001) is a self-report scale designed to screen for the most common *Diagnostic and Statistical Manual of Mental Disorders Fourth Edition* (DSM-IV; American Psychiatric Association, 1994) Axis I disorders encountered in outpatient mental health settings. Development of the PDSQ was based on the following five research and clinical factors that occurred over the past two decades: (1) the need to have standardized instruments to reliably assess published criteria for diagnostic decisions; (2) the development of various self-report questionnaires to diagnosis specific DSM-IV disorders; (3) the importance of diagnosing comorbidity, or the presence of other disorders beyond the primary disorder; (4) the under-recognition of comorbidity in clinical settings due to inadequate measuring instruments; and (5) the need for clinicians to have instruments to administer during the course of routine diagnostic evaluations. Based on these considerations, the PDSQ was designed to assess current and recent symptoms, and be administered and scored within the clinician's office before the diagnostic evaluation (Zimmerman & Mattia, 2001). The scale includes 139 items sampled from the following 15 domains: Major Depressive Disorder (MDD), Dysthymia (DYS), Post-Traumatic Stress Disorder (PTSD), Bulimia Nervosa (BUL), Obsessive Compulsive Disorder (OCD), Panic Disorder (PAN), Mania (MANIA), Psychosis (PSYCH), Agoraphobia (AGOR), Social Phobia (SOC), Alcohol Abuse (ALC), Drug Abuse (DRUG), Generalized Anxiety Disorder (GAD), Somatoform (SOM), and Hypochondriasis (HYP). Scale items are dichotomously scored, with respondents indicating "Yes" (a score of 1) if the item is applicable, or "No" (a score of 0), otherwise.

Purpose

The objective of this study was to investigate the administration of the PDSQ within a CAT environment based on conceptualizing the scale in terms of the bifactor model (Gibbons & Hedeker, 1992). The broader implication of this research is to begin employing CAT methods for multidimensional PRO measurements.

Method

Participants

The data consisted of item responses from 3,791 participants of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial who completed the 139-item PDSQ (Zimmerman & Mattia, 2001) at study entry. All participants met DSM-IV (American Psychiatric Association, 1994) criteria for non-psychotic major depressive disorder (MDD) defined clinically and confirmed by a checklist of symptoms completed by the Clinical Research Coordinator. All participants completed a written informed consent before study entry. Study oversight was provided by Institutional Review Boards at the fourteen participating Regional Centers, relevant clinical sites, and the NIMH Data Safety and Monitoring Board. Detailed descriptions of the STAR*D project are provided by Fava, Rush, Trivedi et al. (2003) and Rush, Fava, Wisniewski et al. (2004).

Instrumentation

The psychometric properties of the PDSQ have been investigated and reported by Zimmerman and Mattia (2001). Internal consistency estimates (Cronbach's coefficient alpha) exceeded 0.80 for all but one subscale, with a mean value of 0.86. Test-retest reliability over a one-week interval exceeded 0.80 for nine subscales, with a mean test-retest correlation of 0.83. Convergent and discriminant validity coefficients indicated that PDSQ subscale scores correlated higher with other scale scores that measured the same construct (e.g., depression) than those that measured other symptoms (e.g., somatoform). A contrasting groups study indicated that subscale scores were statistically higher for patients with the particular diagnosis (e.g., bulimia nervosa) than those without the disorder (e.g., mania). The subscales' diagnostic performance was found to vary according to the cutoff score. Specifically, as the threshold for case identification increased, subscale sensitivity decreased and specificity increased. Furthermore, mean subscale sensitivities of 80%, 85%, and 90% resulted in mean subscale specificities of 78%, 73%, and 66%, with negative predictive values of 95%, 96%, and 97%, respectively.

Receiver operating curves were estimated for each subscale and all areas under the curve were significant (Zimmerman & Mattia, 2001). The PDSQ items are shown in Table 1.

For the data of this study, full scale reliability, based on Cronbach's coefficient alpha, was 0.95. Internal consistency estimates for the PDSQ subscales ranged between 0.83 (Hypochondiasis) and 0.94 (Somatoform), with the exception of the following subscales: Major Depressive Disorder (0.76), Obsessive Compulsive (0.78), Mania (0.70), and Psychosis (0.62).

Data Analysis

Based on previous research (Gibbons, Immekus, & Bock, in press) supporting the fit of the bifactor model over competing IRT models (e.g., unidimensional, simple structure) to describe the PDSQ data, a full-information item bifactor model (Gibbons & Hedeker, 1992) was used for item parameter estimation of the primary dimension and all 15 sub-domains.

Within the bifactor model, item thresholds, primary factor loadings, and domain-specific factor loadings are computed. The threshold describes the point on the underlying primary symptom/impairment dimension (characterized by all items) at which 50% of the sample can be expected to provide a positive response. For example, in the domain of depression, an item with a high threshold (e.g., suicidal ideation) is rated positively only by the most severely depressed patients, whereas an item with a low threshold (e.g., depressed mood) is rated positively by patients with both high and low underlying levels of depression. The primary loading can be interpreted as a factor loading (correlation with the underlying primary dimension) that is appropriate for a dichotomous response measure. The domain-specific factor loading represents the correlation of the item with the underlying domain that the item was sampled from.

POSTSIMB-15 (Weiss, 2007) was used to conduct the post-hoc simulation of the PDSQ full scale and subscale items. Item administration was based on selecting the item with the maximum amount of information for trait (θ) estimation. On the primary dimension, three fixed standard error of measurement (SEM) termination criteria were evaluated: 0.30, 0.25, and 0.20. Separate fixed SEM termination criterions were used for subscales with less than ten items (0.50) and more than ten items (.40). Efficiency of the CAT session was judged on (1) correlation between full and CAT θ estimates ($\hat{\theta}$s); (2) mean/average signed difference between full and CAT $\hat{\theta}$s; (3) mean number of items administered.

Results

Bifactor Analysis

Table 1 reports the thresholds and the primary and group factor loadings for each item based on the bifactor model. The threshold parameters indicate that Item 1 ("Feel sad or depressed past 2 weeks") was endorsed at the lowest level of mental illness. This result is expected since the sample had to have prominent depressive symptoms. By contrast, Item 79 ("Think had special powers") is the item text correct? reported the highest threshold, and was endorsed only at the highest levels of mental illness. This result is consistent with the exclusion of patients with prominent psychotic features. In terms of domains, DYS and GAD had low thresholds, whereas MANIA, PSYCH, ALC and DRUG had uniformly high thresholds. The low thresholds for DYS and GAD are expected since DYS is often a concomitant of MDD, as is GAD.

Table 1. PDSQ Items, Thresholds, and Factor Loadings From the Bifactor Model

Domain/			Primary	Domain
Item	Question	Threshold	Loading	Loading
MDD: Major Depression				
1	Feel sad or depressed past 2 weeks	-2.15	0.17	0.52
2	Sad/depressed every day past 2 weeks	-0.72	0.26	0.43
3	Less pleasure from things 2 weeks	-1.13	0.17	0.39
4	Less interest in most activities 2 weeks	-1.13	0.20	0.35
5	Appetite significantly lower 2 weeks	0.18	0.19	0.13
6	Appetite significantly greater 2 weeks	0.66	0.10	0.03
7	Sleep at least 1-2 hours less 2 weeks	-0.13	0.25	0.08
8	Sleep at least 1-2 hours more 2 weeks	0.70	-0.04	0.11
9	Feel jumpy and restless 2 weeks	0.15	0.39	0.09
10	Tired nearly every day past 2 weeks	-1.15	0.18	0.24
11	Feel guilty about things 2 weeks	-0.60	0.36	0.33
12	Negative thoughts about self 2 weeks	-0.68	0.30	0.49
13	Feel like failure past 2 weeks	-0.38	0.34	0.54
14	Problems concentrating every day past 2 weeks	-0.83	0.34	0.22
15	Decision making more difficult 2 weeks	-0.51	0.33	0.28
16	Think of dying in passive ways 2 weeks	0.22	0.26	0.73
17	Wish you were dead 2 weeks	0.56	0.19	0.90
18	Think you’re better off dead 2 weeks	0.32	0.22	0.85
19	Thoughts of suicide past 2 weeks	0.26	0.19	0.75
20	Seriously consider taking life 2 weeks	1.23	0.23	0.77
21	Think specific way to take life 2 weeks	0.91	0.17	0.74
DYS: Dysthymia				
22	Feel sad/down most days past 2 years	-0.34	0.35	0.79
23	Poor appetite/overeat most days 2 years	-0.03	0.36	0.58
24	Not sleep enough/too much sleep 2 years	-0.49	0.35	0.68
25	Tired most days past 2 years	-0.48	0.32	0.80
26	Problem concentrating/making decisions 2 years	-0.18	0.41	0.71
27	Low self-esteem most days 2 years	-0.53	0.41	0.66
28	Feel hopeless about future 2 years	-0.23	0.42	0.64
PTSD: Post-Traumatic Stress				
29	Ever experienced traumatic event	0.09	0.31	0.48
30	Ever witnessed traumatic event	0.36	0.29	0.40
31	Thoughts of trauma pop into mind	0.19	0.41	0.75
32	Upset because thinking of trauma	0.39	0.43	0.75
33	Bothered by memory/dreams of trauma	0.24	0.45	0.78
34	Reminders of trauma cause distress	0.29	0.48	0.76
35	Block out thought/feeling of trauma	0.12	0.45	0.74
35	Block out thought/feeling of trauma	0.12	0.45	0.74
36	Avoid activities remind of trauma	0.46	0.48	0.67
37	Flashbacks of traumatic event	0.63	0.50	0.64
38	Reminders make you shake	0.70	0.57	0.59
39	Feel distant because of trauma	0.52	0.46	0.73
40	Feel numb because of trauma	0.54	0.43	0.70
41	Give up goals because of trauma	0.78	0.46	0.63
42	Keep guard up because of trauma	0.24	0.45	0.70
43	Jumpy because of a trauma	0.67	0.52	0.62

Table 1 (cont'd). PDSQ Items, Thresholds, and Factor Loadings from the Bifactor Model

Domain/ Item	Question	Threshold	Primary Loading	Domain Loading
BUL: Bulimia Nervosa				
44	Often go on eating binges	0.47	0.33	0.86
45	Can't control how much you eat	0.69	0.31	0.85
46	Eat so much uncomfortably full	0.36	0.28	0.87
47	Eat a lot when not hungry	0.44	0.24	0.88
48	Eat alone because embarrassed	0.90	0.28	0.82
49	Feel disgusted after overeating	0.47	0.27	0.90
50	Upset with self because of binges	0.56	0.28	0.89
51	Strict diets, exercise excessively	1.20	0.30	0.55
52	Force self to vomit	1.61	0.30	0.55
53	Weight most important thing	0.10	0.21	0.51
OCD: Obsessive Compulsive Disorder				
54	Worry about dirt, germs	1.25	0.46	0.44
55	Worry something you forgot	0.53	0.55	0.41
56	Worry you'd act/speak violently	0.55	0.60	0.30
57	Compelled to do things over and over	1.21	0.50	0.66
58	Do things over that interfered	0.98	0.47	0.66
59	Wash and clean excessively	1.22	0.47	0.60
60	Excessively check and do things over	0.91	0.51	0.68
61	Count things obsessively/excessively	1.31	0.44	0.55
PAN: Panic				
62	Scared because heat beating fast	0.53	0.45	0.72
63	Scared because short of breath	0.67	0.48	0.71
64	Scared because shaky or faint	0.57	0.52	0.66
65	Anxiety attacks for no reason	0.21	0.59	0.49
66	Anxiety attacks, think will go crazy	0.41	0.64	0.45
67	Anxious attacks with 3 or more symptoms	0.30	0.60	0.62
68	Worry about having anxiety attacks	0.69	0.63	0.44
69	Anxiety attacks caused avoid situations	0.45	0.64	0.30
70	Feel excessively cheerful/happy	1.11	0.14	0.84
MANIA				
71	Feel extremely self-confident	1.18	0.13	0.87
72	So much energy, need less sleep	1.40	0.16	0.84
73	Talk more than usual	1.05	0.35	0.60
74	Thought could do everything	1.01	0.24	0.62
75	Do impulsive things	1.03	0.32	0.49
PSYCH: Psychosis				
76	People tell imagination	1.20	0.49	0.50
77	Convinced others spying	0.85	0.55	0.55
78	Think danger because someone plotting	1.55	0.56	0.48
79	Think had special powers	2.00	0.41	0.51
80	Think some force controlled	1.91	0.49	0.53
81	See/hear things other people didn't	1.64	0.47	0.49

Table 1 (cont'd). PDSQ Items, Thresholds, and Factor Loadings From the Bifactor Model

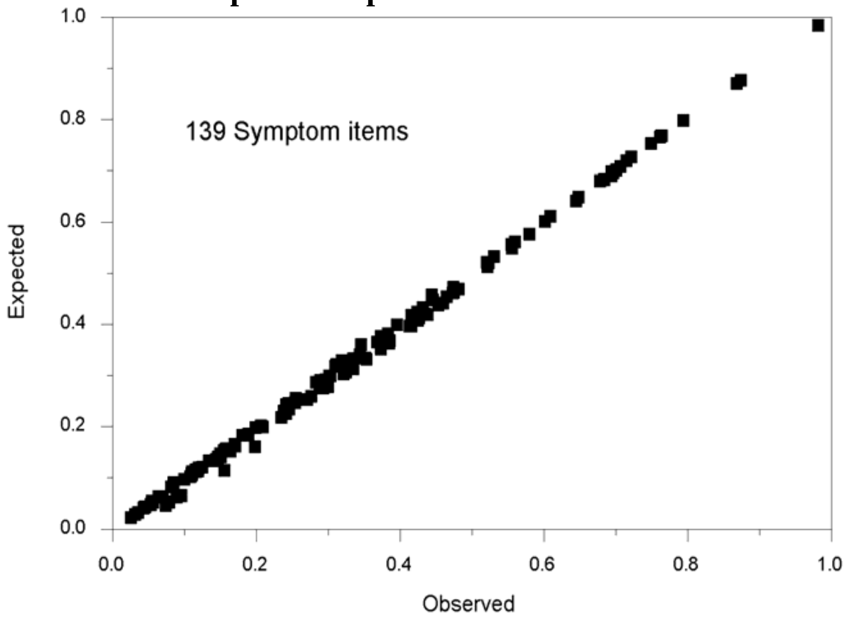
Domain/			Primary	Domain
Item	Question	Threshold	Loading	Loading
AGOR: Agoraphobia				
82	Avoid situation because afraid of anxiety attack	0.69	0.64	0.35
83	Anxious going far away from home	1.06	0.59	0.53
84	Anxious being in crowded places	0.54	0.64	0.61
85	Anxious standing in long likes	0.80	0.64	0.54
86	Anxious being on bridge or in tunnel	1.11	0.49	0.50
87	Anxious traveling in bus, train, plane	1.09	0.49	0.57
88	Anxious driving/riding in a car	1.12	0.51	0.49
89	Anxious being home along	1.01	0.52	0.28
90	Anxious being in open spaces	1.68	0.59	0.50
91	Get anxious as soon as in situation	0.59	0.64	0.61
92	Avoid situation because made you anxious	0.44	0.63	0.62
SOC: Social Phobia				
93	Worry about embarrassing self	0.08	0.52	0.64
94	Worry you'd say something stupid	-0.03	0.50	0.66
95	Nervous when people pay attention	-0.12	0.46	0.69
96	Nervous in social situations	0.10	0.52	0.65
97	Avoid situations because afraid embarrass self	0.34	0.57	0.64
98	Worry public speaking	0.12	0.39	0.64
99	Worry eating in front of others	0.85	0.47	0.49
100	Worry using public restrooms	1.27	0.45	0.34
101	Worry writing in front of others	1.08	0.43	0.39
102	Worry saying something stupid	0.15	0.46	0.76
103	Worry asking questions around others	0.27	0.44	0.72
104	Worry business meetings	0.73	0.37	0.58
105	Worry parties/social gatherings	0.38	0.46	0.74
106	Get anxious as soon as in situation	0.24	0.53	0.66
107	Avoid situations because made you anxious	0.20	0.54	0.62
ALC: Alcohol Abuse				
108	Think drink too much	1.21	0.08	0.92
109	Family say drink too much	1.51	0.18	0.86
110	Doctor/friends say drink too much	1.68	0.13	0.88
111	Think about cutting down on drinking	0.99	0.05	0.91
112	Think had alcohol problem	1.55	0.11	0.83
113	Problem with marriage because of drinking	1.64	0.19	0.84
DRUG: Drug Abuse				
114	Think using drugs too much	1.54	0.19	0.93
115	Family say use drugs too much	1.73	0.28	0.85
116	Doctor/friends say use drugs too much	1.85	0.24	0.89
117	Think about cutting down on drug use	1.34	0.26	0.89
118	Think had a drug problem	1.74	0.17	0.89
119	Problem with marriage because of drug use	1.71	0.29	0.86

Table 1 (cont'd). PDSQ Items, Thresholds, and Factor Loadings From the Bifactor Model

Domain/ Item	Question	Threshold	Primary Loading	Domain Loading
GAD: Generalized Anxiety				
120	Nervous person most days	-0.05	0.56	0.43
121	Worry bad things happened	-0.08	0.59	0.36
122	Worry about things shouldn't	-0.28	0.52	0.42
123	Worry daily	-0.48	0.52	0.67
124	Feel restless because worrying	-0.53	0.57	0.66
125	Problem falling asleep because anxiety	-0.45	0.49	0.47
126	Tension in muscles because anxiety	-0.57	0.53	0.42
127	Trouble concentrating because worrying	-0.73	0.58	0.57
128	Snappy/irritable because worrying	-0.73	0.45	0.42
129	Hard to control worrying	-0.51	0.54	0.69
SOM: Somatoform				
130	Had a lot of stomach problems	0.19	0.35	0.46
131	Bothered by aches/pains	-0.14	0.41	0.53
132	Get sick more than most people	0.85	0.35	0.75
133	Health been poor most of life	1.26	0.35	0.64
134	Doc not able to find cause for sick	1.02	0.37	0.47
HYPO: Hypochondiasis				
135	Worry might have serious illness	0.43	0.44	0.83
136	Hard to stop worrying about illness	0.74	0.49	0.83
137	Doctor said didn't have illness	1.13	0.48	0.64
138	Worry illness, interfere with activities	1.11	0.55	0.68
139	Visit doctor much because worried about illness	1.26	0.46	0.65

In terms of loadings on the primary dimension (interpreted as factor loadings on the overall mental illness dimension), the ALC and DRUG domains had the lowest loadings on the primary mental illness dimension (on average about 0.20), whereas PTSD, OCD, PAN, PSYCH, AGOR, SOC, GAD, and HYPO had the highest (on average about 0.50). Interestingly, the MDD items had lower loadings on the primary dimension, indicating that the variance of these items was not accounted for as much by the primary dimension. Nevertheless, given the diversity of the PDSQ subscales, primary loadings cannot be considered negligible, as they were predominantly in the low to moderate range (0.20 to 0.60). Therefore, the magnitude of the loadings on the primary dimension suggests the role it plays in accounting for the relationship among the PDSQ items. Figure 1 presents observed and predicted response proportions for the fifteen sub-domain bi-factor model, and it illustrates excellent fit of the model to the observed 139 symptom response proportions (*r* = 0.99).

Figure 1. Comparison of Observed and Expected Symptom Response Proportions for the Bifactor Model



Post-Hoc Simulation

POSTSIMB-15 (Weiss, 2007) was used to conduct the post-hoc simulation administration of the PDSQ on the primary dimension and subsequently on each secondary dimension. Specifically, the program estimated each respondent's θ (impairment level) on the primary dimension. Then, using the responses to the items for a given secondary dimension that had been answered by the respondent in estimating their θ on the primary dimension, in conjunction with the item discriminations on the secondary dimension, a starting $\hat{\theta}$ was computed for each successive secondary dimension (see Weiss & Gibbons, 2007, for details).

Table 2 reports the post-hoc simulation results for the PDSQ general factor. As shown, the correlation between full and CAT $\hat{\theta}$s increased with a more stringent SEM termination criterion. Specifically, full and CAT $\hat{\theta}$s were moderately correlated for a SEM termination criteria of 0.30 (r = 0.69), and more strongly correlated for a termination of 0.20 (r = 0.83). As shown in the table, the termination of 0.20 was aligned with the mean full scale SEM estimate. The mean full scale $\hat{\theta}$ was -1.94, indicating that on average the respondents reported generally low levels of psychiatric impairment. For the SEM termination of 0.20, the average number of items administered was 71.70 (*SD* = 36.53), a 52% reduction compared to the full scale. However, the full and CAT $\hat{\theta}$ correlation of 0.83 was lower than the desired 0.90 criterion.

As shown in Figure 2, a bivariate plot of the full and CAT $\hat{\theta}$s revealed an underestimation of CAT θs for many respondents. Examination of the item response vectors of these respondents indicated that they endorsed only a small number of items across the 15 PDSQ subscales.

Table 2. Results of Post-Hoc Simulation Study of PDSQ for General Factor (N = 3,791)

SEM Termination	Corre-lation	No. of Items		θ	$\hat{\theta}$ Descriptive Statistics			SEM Descriptive Statistics		
		Mean	SD		Mean	SD	Range	Mean	SD	Range
				Full	−1.94	0.96	−3.96 - 1.08	0.19	.028	0.09 - 0.25
0.30	0.69	31.53	17.37	CAT	−1.03	2.07	−3.58 - 1.57	0.30	.003	0.28 - 0.30
0.25	0.74	44.87	22.75	CAT	−1.63	1.91	−3.71 - 1.08	0.25	.005	0.23 - 0.25
0.20	0.83	71.70	36.53	CAT	−1.82	1.56	−3.80 - 1.08	0.20	.008	0.19 - 0.25

Based on the preliminary nature of this study, follow-up post-hoc simulation administrations of the primary dimension were conducted by excluding respondents who had full and CAT $\hat{\theta}$s that had an absolute difference greater than 1.50 and 1.00, respectively. The objective of these analyses was to better understand the nature of the functioning of adaptively administering the PDSQ based on the bifactor model.

Figure 2. Plot of Full vs. CAT θ Estimates

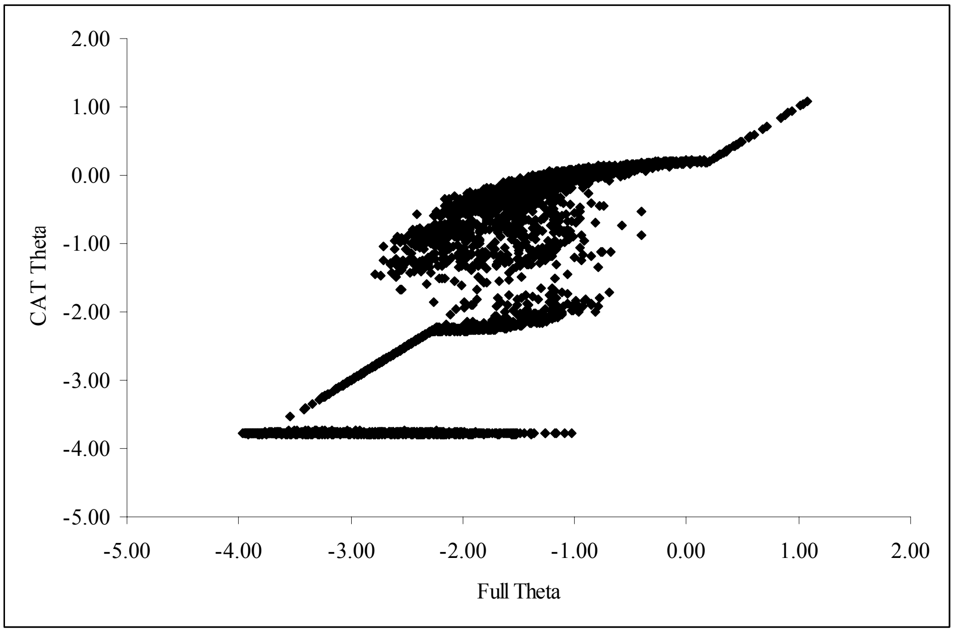


Table 3 provides the results of the post-hoc simulation of the primary dimension, excluding respondents with full and CAT $\hat{\theta}$s greater than 1.50. As shown, full and CAT $\hat{\theta}$s were more highly correlated. The full scale SEM was 0.191, so the SEM termination of 0.20 seems adequate for these data (similar to previous analyses). Based on the SEM termination of 0.20, full and CAT $\hat{\theta}$ correlations approached the criterion of 0.90, with a value of 0.87. The mean number of items administered was 73.84 (SD= 36.89).

Table 3. Results Based on Excluding Examinees With CAT and Full-Scale $\hat{\theta}$s That Differed More Than an Absolute Value of 1.50 on the General Factor (*N* = 3,446)

SEM	Corre-	No. of Items			$\hat{\theta}$ Descriptive Statistics			SEM Descriptive Statistics		
Termination	lation	Mean	SD	θ	Mean	SD	Range	Mean	SD	Range
				Full	–1.94	1.00	–3.96 - 1.08	0.19	0.03	0.09 - 0.25
0.30	0.79	32.39	18.35	CAT	–1.40	2.09	–3.58 - 1.57	0.30	0.01	0.28 - 0.30
0.25	0.77	46.12	23.55	CAT	–1.57	1.90	–3.71 - 1.08	0.25	0.01	0.23 - 0.25
0.20	0.87	73.84	36.89	CAT	–1.79	1.56	–3.79 - 1.08	0.20	0.01	0.19 - 0.25

Table 4 reports the results of the post-hoc simulation administration of the primary dimension, excluding respondents with full and CAT $\hat{\theta}$s with an absolute difference greater than 1.00. Across SEM termination conditions, full and CAT $\hat{\theta}$s correlated above 0.80. Based on a SEM termination of 0.20, the full and CAT $\hat{\theta}$ correlation exceeded 0.90. This criterion also resulted in mean CAT $\hat{\theta}$s that approached full scale $\hat{\theta}$s, with a substantial reduction in number of administered items (M = 78.88, SD = 39.32).

Table 4. Results Based on Excluding Examinees With CAT and Full-Scale $\hat{\theta}$s That Differed More Than an Absolute Value of 1.00 on the General Factor (*N* = 2,327)

SEM	Corre-	No. of Items			$\hat{\theta}$ Descriptive Statistics			SEM Descriptive Statistics		
Termination	lation	Mean	SD	θ	Mean	SD	Range	Mean	SD	Range
				Full	–1.99	0.03	–3.96 - 1.08	0.19	0.03	0.05 - 0.25
0.30	0.85	33.69	20.73	CAT	–1.52	2.09	–3.58 - 1.57	0.30	0.01	0.28 - 0.30
0.25	0.85	48.37	25.83	CAT	–1.68	1.92	–3.71 - 1.08	0.25	0.01	0.23 - 0.25
0.20	0.93	78.88	39.32	CAT	–1.86	1.58	–0.79 - 1.08	0.20	0.01	0.19 - 0.25

Table 5 shows the results of the post-hoc simulation of the 15 PDSQ secondary dimensions. Inspection of the table indicates that full and CAT $\hat{\theta}$s were highly correlated ($r \geq 0.95$). The mean number of items administered for dimensions with greater than ten items was generally only a few items less than the total number of items comprising the subscale. Full scale SEM values were similar to the SEM termination criterion selected prior to data analysis. That is, they were generally equal to or less than 0.50. The proportion of reused items indicates the number of items administered in the post-hoc simulation during the administration of the primary dimension and again for the secondary dimension. Secondary dimensions with the largest proportion of reused items were AGOR, HYPO, PTSD, and OCD. Those with the least proportion of reused items were MDD, MANIA, and ALC. Inspection of Table 1 indicates that subscales comprised of higher loadings on the primary dimension (e.g., AGOR, HYPO) resulted in a higher proportion of reused items than those with low loadings on the primary dimension (e.g., MDD).

Table 5. Results of Post-Hoc Simulation Study of PDSQ for Secondary Factors

	No. of	No. of CAT Items		Corre-	Full (F) and CAT (C) $\hat{\theta}$s				Full (F) and CAT (C) SEMs				Proportion Reused	
Scale	Items	Mean	SD	lation		Mean	SD	Range		Mean	SD	Range	Mean	SD
1	21	16.85	6.71	0.98	F	–1.06	1.78	–3.47 - 3.10	F	0.54	0.15	0.34 - 1.26	0.32	0.47
					C	–1.30	2.00	–3.47 - 3.10	C	0.52	0.16	0.34 - 1.26		
2	7	6.46	1.06	0.99	F	1.10	0.97	–0.92 - 2.09	F	0.48	0.09	0.39 - 0.69	0.55	0.50
					C	1.08	0.98	–0.92 - 2.09	C	0.50	0.08	0.39 - 0.69		
3	15	13.67	2.41	0.99	F	–1.19	1.08	–2.41 - 1.13	F	0.39	0.10	0.26 - 0.64	0.68	0.47
					C	–1.19	1.09	–2.41 - 1.13	C	0.40	0.09	0.26 - 0.64		
4	10	6.60	2.86	0.90	F	–2.41	1.02	–3.23 - 0.49	F	0.39	0.10	0.25 - 0.81	0.67	0.47
					C	–2.12	0.97	–3.23 - 0.49	C	0.42	0.09	0.25 - 0.81		
5	8	7.06	0.89	0.98	F	–2.56	0.68	–2.99 - 0.57	F	0.48	0.06	0.45 - 0.83	0.69	0.46
					C	–2.48	0.67	–2.99 - 0.57	C	0.48	0.05	0.45 - 0.83		
6	9	8.54	1.18	0.99	F	–1.25	0.96	–2.14 - 1.00	F	0.50	0.08	0.41 - 0.71	0.66	0.47
					C	–1.25	0.96	–2.14 - 1.00	C	0.50	0.08	0.41 - 0.71		
7	5	4.49	0.50	0.97	F	–3.43	0.60	–3.66 - 0.08	F	0.35	0.14	0.28 - 0.96	0.28	0.49
					C	–3.28	0.56	–3.66 - 0.08	C	0.42	0.12	0.28 - 0.96		
8	6	6.00	0.00	1.00		–2.91	0.45	–3.13 - 0.30		0.54	0.05	0.51 - 0.87	0.54	0.49
9	11	10.05	1.75	0.97	F	–2.06	0.93	–2.85 - 0.85	F	0.47	0.05	0.39 - 0.72	0.79	0.40
					C	–1.95	1.03	–2.85 - 0.85	C	0.48	0.04	0.39 - 0.72		
10	15	13.52	2.26	0.99	F	–0.83	1.15	–2.39 - 1.38	F	0.41	0.11	0.29 - 0.59	0.68	0.47
					C	–0.79	1.17	–2.39 - 1.38	C	0.42	0.10	0.29 - 0.59		
11	6	3.04	1.63	0.95	F	–3.74	0.68	–3.93 - 0.01	F	0.17	0.17	0.11 - 0.99	0.23	0.42
					C	–3.41	0.62	–3.93 - 0.01	C	0.38	0.16	0.11 - 0.99		
12	6	5.70	1.78	0.95	F	–3.86	0.50	–3.95 - 0.00	F	0.12	0.13	0.09 - 0.99	0.46	0.50
					C	–3.69	0.49	–3.95 - 0.00	C	0.23	0.15	0.09 - 0.99		
13	10	9.00	1.99	0.99	F	1.19	0.98	–1.18 - 2.19	F	0.50	0.08	0.41 - 0.71	0.59	0.49
					C	1.17	0.98	–1.18 - 2.19	C	0.51	0.07	0.41 - 0.71		
14	5	5.00	0.00	1.00		–1.97	0.86	–2.73 - 0.75		0.64	0.07	0.57 - 0.82	0.44	0.50
15	5	4.35	1.06	0.96	F	–2.03	0.83	–2.58 - 0.34	F	0.47	0.09	0.33 - 0.80	0.75	0.44
					C	–1.89	0.81	–2.58 - 0.34	C	0.47	0.09	0.33 - 0.80		

Discussion

This study represents a preliminary investigation into the potential contributions of the bifactor IRT model and CAT in PRO measurements. To date, CAT applications have been based on traditional unidimensional IRT models. Ubiquitous to these models is that a single latent trait (e.g., depression, anxiety) underlies item responses. However, PRO measures are typically multidimensional in nature. Consequently, the multifaceted nature of these scales thus limits the direct application of traditional IRT models to PRO scale data. Although the use of unidimensional IRT models and CAT for the delivery of PRO measurements has been investigated (e.g., Fliege, Becker, Walter, et al., 2005), research is needed to study the use of IRT models (e.g., the bifactor model) that may better capture the dimensionality present in PRO data.

To address this issue, a post-hoc simulation administration of the PDSQ within a CAT framework was conducted based on conceptualizing the scale's factor structure in terms of the bifactor model. The full-information item bifactor model represents a recently derived multidimensional, confirmatory-based IRT model. The premise of the model is that each item is related to a primary dimension and only one secondary dimension. Unlike unrestricted full-information item factor analysis (Bock, Gibbons, & Muraki, 1988), the bifactor model can be used to model data with a large number of sub-domains because the equation always reduces to a two-dimensional integral (Gibbons & Hedeker, 1992). Furthermore, it can be used as a basis to provide a stringent test of the number of dimensions underlying a scale's data by comparing the fit (testing the difference between log likelihood values) between competing IRT models (e.g., unidimensional model vs. bifactor model). Within this study, the bifactor model was found to adequately describe the PDSQ scale data, with each of the 15 symptom sub-domains improving overall model-data fit.

The post-hoc simulation administration of the PDSQ indicated that there are benefits of incorporating the bifactor model and CAT into the delivery of PRO measurements. First, there was approximately a 50% reduction in administered items on the primary dimension, although less pronounced on the secondary dimensions. Given that a set of clinical scales may be administered for diagnostic purposes, by obtaining comparable measurement precision through the use of minimal items via CAT represents direct cost savings (e.g., time). Second, as is often desired in clinical settings, respondents can be scored on the primary and the secondary dimensions. Therefore, primary and secondary trait scores can be used by practitioners for various inferential decision-making purposes (e.g., diagnosis, treatment efficacy). Although this study represented an initial step in the use of the bifactor model and CAT in PRO measurements, it suggests there may be a more direct approach to administering multidimensional scales via computer as opposed to administering separate unidimensional scales.

Nevertheless, the results of this study point to areas of consideration. First, inspection of the bifactor results indicates that primary dimension loadings varied depending on sub-domain content. For example, items with the highest loading fell within the AGOR subscale, whereas the lowest loadings were associated with ALC items. Within the context of CAT, items with the highest loadings on the measured dimension will be selected for administration before items with low loadings, due to the amount of information they provide in θ estimation. Consequently, in certain instances, a potentially important sub-domain may not be included in the administration of the primary dimension items due to their low relationship with the general factor. This may have occurred in the present study due to the moderate correlation observed between full and

CAT $\hat{\theta}$s on the primary dimension, as many respondents' CAT $\hat{\theta}$s were underestimated. This issue may be resolved through the use of scales that report moderate to high loadings on both the primary and secondary dimensions. Reise, Morizot, and Hays (in press) discuss when researchers should consider using the bifactor model over unidimensional IRT models.

The results of this study also indicate areas of future research. First, research needs to investigate how strong primary dimension loadings should be in the bifactor model for CAT scale administration. This relates to the issue of content balancing and deals with the instance when a scale's sub-domains span across a range of content and the collective item set demonstrates low to high loadings on the primary dimension, as is observed in the PDSQ. Another issue is the size of the item bank on the sub-domain, or "group" factor. Given that PRO measures are typically comprised of subscales that may have less than 15 items, research needs to address how these items should be administered within CAT. In these data, the proportion of reused sub-domain items generally exceeded 40%. Another issue is the effect of a lack of item parameter invariance (i.e., differential item functioning) across respondent groups (e.g., depressed vs. non-depressed respondents). Each of these empirical questions relates to the efficacy of using the bifactor model and CAT in PRO measurement and represent areas for future research.

References

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders, 4th edition.* Washington, DC: Author.

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

Bock R. D., Gibbons R. D., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12,* 261-280.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries*. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory: Research Report 77-6.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*, 189-225.

Fava, M., Rush, A. J., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sackeim, H. A., Quitkin, F. M., Wisniewski, S., Lavori, P. W., Rosenbaum, J. F. & Kupfer, D. J. (2003). Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study. *Psychiatric Clinics of North America*, *26,* 457-494.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of life Research, 14*, 2277-2291.

Gibbons, R. D., & Hedeker D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423-436.

Gibbons, R. D., Immekus, J. C., & Bock, R. D. (in press). *Didactic workbook: The added value of multidimensional IRT models*.

Holzinger, K. J., & Swineford, F. (1937). The bifactor method. *Psychometrika*, *2*, 41-54.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations, *Psychometrika, 54,* 557-585.

Reise, S. P., Morizot, J., & Hays, R. D. (in press). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., Kashner, T. M., Kupfer, D. J., Rosenbaum, J. F., Alpert, J., Stewart, J. W., McGrath, P. J., Biggs, M. M., Shores-Wilson, K., Lebowitz, B. D., Ritz, L., & Niederehe, G. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): Rationale and design. *Controlled Clinical Trials, 25,*119-142.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, *53,* 774-789.

Weiss, D. J. (2007). POSTSIMB-15 [Computer Program]. St. Paul MN: Assessment Systems Corporation.

Weiss, D. J. & Gibbons, R.D. (2007, June). *CAT with the bifactor model.* Paper presented at the 2007 GMAC Conference on Computerized Adaptive Testing. Minneapolis MN.

Weiss, D. J., & Kingsbury, G. G. (1984), Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21,* 361-375.

Zimmerman, M., & Mattia, J. I. (2001). A self-report scale to help make psychiatric diagnoses: the Psychiatric Diagnostic Screening Questionnaire. *Archives of General Psychiatry*, *58,* 787-794.