Graduate Management Admission Council®

Computerized Adaptive Testing With the Bifactor Model

David J. Weiss
University of Minnesota

and

Robert D. Gibbons
Center for Health Statistics
University of Illinois at Chicago

Presented at the New CAT Models session, June 8, 2007

Abstract

An algorithm designed to implement CAT with the dichotomous bifactor model is described. Performance of the algorithm was evaluated with several datasets from a 615-item personality instrument that scored on a general scale and four content scales. Post-hoc simulation, including cross-validation, and live testing bifactor CAT data were analyzed in terms of reductions in test length for each scale, correlations with trait estimates from all items in each scale, bias, and accuracy. Results showed very substantial reductions in scale and overall test length while maintaining correlations with full-scale scores above .90. For the general scale, mean test length reductions of about 95% were observed in both post-hoc simulation and live testing; only about 25 to 30 items were required, on average, to recover scale scores with a correlation above .90. Mean reductions of 68% to 90% were observed for the content scales. Across all scales combined, the bifactor CAT algorithm reduced test length by an average of about 80% and resulted in an actual testing time mean decrease of approximately 93 minutes (82%).

Acknowledgments

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC® and by NIMH grants R01-MH66302 and R01-MH30915, Robert D. Gibbons, Principal Investigator. The contributions of David J. Kupfer, Ellen Frank, Andrea Fagiolini, Victoria J. Grochocinski, Dulal K. Bhaumik, Angela Stover, R. Darrell Bock, and Jason C. Immekus, in project planning and data collection are gratefully acknowledged.

Citation

Weiss, D. J. & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. 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

David J. Weiss, N660 Elliott Hall, University of Minnesota, Minneapolis MN 55455, U.S.A. djweiss@umn.edu

Computerized Adaptive Testing With the Bifactor Model

Most applications of computerized adaptive testing (CAT) have used unidimensional item response theory (IRT) models. These models are appropriate for many psychological variables that account for individual differences along a single psychological dimension. These include ability type variables as well as many personality variables.

Some psychological variables, however, are multidimensional. These also include ability types of variables as well as personality variables. Intelligence, for example, can be characterized as a single unidimensional construct along which individuals vary, or it can be described as a hierarchical construct that includes specific kinds of cognitive abilities (verbal ability, reasoning, quantitative ability) that have a common component that might be referred to as "general" intelligence. Similarly, personality variables such as depression can be viewed as single construct, or as a "general" construct that has different components such as sleep disturbances, lethargy, and loss of appetite. For either intelligence or depression, or other variables that have a two-level structure, the applied measurement objective is to measure each individual examinee on the "general" component of each variable as well as each of the lower level more specific components.

When a psychological trait is multidimensional, there are two general approaches to modeling it with IRT. One is to apply multidimensional IRT models (Bock and Aitkin, 1981; Bock Gibbons and Muraki, 1988; Reckase, 1985; Reckase & McKinley, 1991), which result in item parameters for each item that describe the item's contribution to each of the underlying traits that account for the item responses. Application of these multidimensional IRT is similar to implementing a exploratory factor analysis of the item response data and, in essence, results in distributing the variance of each item among the factors that account for the data. Once these item parameters are estimated, CAT could proceed by applying multidimensional CAT algorithms in conjunction with the multidimensional item parameters (e.g., Segall, 1966, 2000; van der Linden, 1999). One problem with this approach is that multidimensional IRT parameter estimation methods have not been thoroughly studied and might require very large sample sizes. A second problem is that the resulting structure for the domain under investigation might not result in trait (θ) estimates that provide the kind of information that is useful for applied purposes, such as measurements on the lower level components that are assumed to underlie the "general" variable..

A plausible alternative factor structure is the "bifactor" model (Holzinger & Swineford, 1937). The bifactor solution constrains each item to have a non-zero loading on the primary dimension (e.g., depression) and a secondary loading on no more than one of the domain content (lower level) factors (e.g., sleep disturbance). The bifactor structure is plausible in mental health measurement, where symptom items that are related to a primary dimension of interest are often selected from underlying measurement sub-domains. It is also plausible for measuring intelligence and many other psychological constructs that have a two-level structure of a "general" variable and content-specific sub-domains

Gibbons and Hedeker (1992) derived a bifactor model for binary response data, and Gibbons, Bock, Hedeker et. al. (2007) extended it for analysis of graded response data. Their estimation method permits the items to be sampled from any number of sub-domains. In the present context, the advantage of the bifactor model is that it yields an overall or "general" measure that can be

the focus of CAT, as well as measurement on the underlying sub-domains. In contrast to the usual multidimensional IRT approach, the bifactor model is a confirmatory model and it retains that hypothesized content structure that is assumed to underly the construct being measured.

Table 1 presents a schematic bifactor structure. An "X" in the table indicates a high loading for the variable on a factor. As Table 1 shows, the general factor is defined by the fact that all items have high loadings on it. Each of the group factors has high loadings on a subset of the items, and each item is constrained to have a high loading on only group (or content) factor. This structure is specified in advance by the researcher. The model is then fit to the data using a confirmatory factor analysis procedure using a computer program such as TESTFACT (Wood, Wilson, Gibbons, Schilling, Muraki, & Bock, 2002). If the model is a good representation of the data, the result is a set of item thresholds and factor loadings that can be converted into IRT item parameters. For binary response data, there is a single threshold for each item, which is converted into the item difficulty (b_i) parameter. Because each item loads on the general factor and at most one content factor, there is also a discrimination (a_i) parameter for each item on the general factor and one content factor.

Table 1
An Illustrative Bifactor Structure:
An "X" Indicates a High Factor Loading

Item	General Factor	Group Factor			
		1	2	3	4
1	X	X			
2	X	X			
3	X	X			
4	X	X			
5	X	X			
6	X		X		
7	X		X		
8	X		X		
9	X		X		
10	X			X	
11	X			X	
12	X			X	
13	X				X
14	X				X
15	X				X
16	X				X
17	X				X
18	X				X

Because the bifactor structure is useful for characterizing a number of psychological constructs and existing multidimensional CAT procedures are inappropriate for use with the bifactor model, a CAT algorithm was developed to use with this IRT model. This paper

describes that algorithm and reports on its utility in both post-hoc simulation and live-testing CAT.

The Bifactor CAT Algorithm

The bifactor CAT algorithm consisted of the following steps:

1. Fit the bifactor model to an appropriately structured set of item responses.
2. Convert the intercept parameter estimate for each item (γ_i) from the bifactor solution to the b_i parameter of the 2-parameter logistic IRT model by

   $$b_i = -\gamma_i / a_{iG} \quad (1)$$

   where a_{iG} is the item discrimination parameter for item *i* on the General factor, and the two-parameter logistic model is given by

   $$P(u_{ij} = 1 | a_i, b_i, \theta_j) = \frac{1}{1 + \exp[-Da_i(\theta_j - b_i)]} \quad (2)$$

   with u_{ij} = the response of examinee *j* to item *i* scored 1 for a keyed response and 0 otherwise,

   θ_j = the trait level of examinee *j*, and

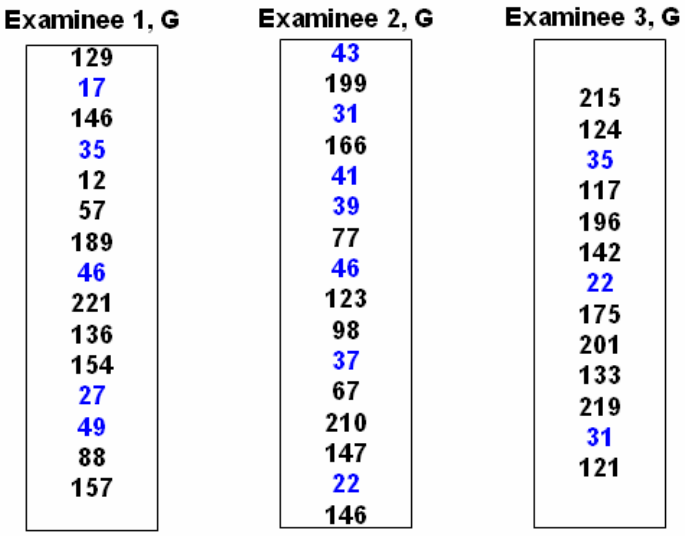
   D = 1.7.
3. Implement CAT on the General scale for each examinee. Each CAT was begun with an initial θ estimate of 0.0, items were selected by maximum information, θ was estimated using Bayesian modal estimation (MAP, or maximum a posteriori), and the CAT was terminated using a fixed standard error of the θ estimate (SEM), allowing the number of items to vary across examinees.
4. Identify, for Content Scale 1, those items that were administered on the examinee's General factor CAT. These items will vary among examinees based on their estimated θ level on the General factor.
5. Using the discrimination parameters from the bifactor solution for Content Scale 1, compute a θ estimate from these items, and use it as the CAT starting θ estimate for Content Scale 1.
6. Implement CAT for Content Scale 1 using its discrimination parameters and an appropriate termination criterion. The content scale CATs used the same set of CAT options as the General scale CAT with two exceptions (1) an examinee's CAT on a Content scale used a variable entry θ estimate, based on Steps 4 and 5; and (2) although a fixed SEM was used to terminate the content scales, the values of the SEM varied among the scales. These values were based on the first set of post-hoc simulations in which post-hoc simulation correlations were sought between CAT $\hat{\theta}$ and full-scale $\hat{\theta}$ of .90 or greater. The SEMs values associated with these results were used to terminate content scale CATs for subsequent post-hoc simulations and live CATs.

7. Repeat Steps 5 -7 for each additional content scale:
    a. Identify administered items from the General Scale for a given content scale.
    b. Estimate θ for the content scale from those items and their scale discrimination parameters.
    c. Implement CAT for that scale.

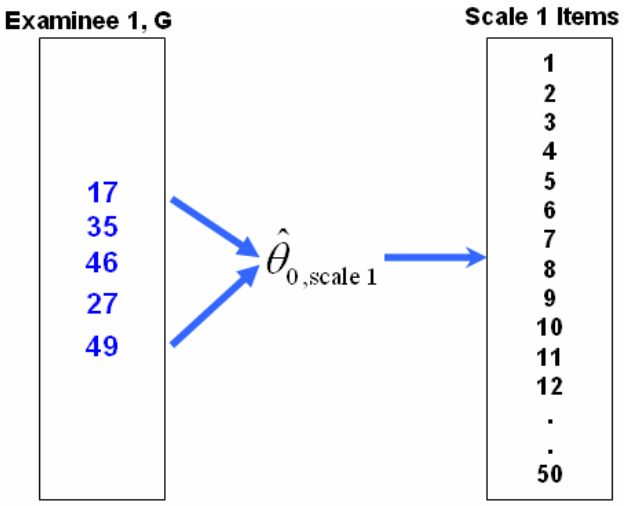
Figure 1 illustrates how items administered on the General (G) scale were used in the bifactor CAT algorithm. Hypothetical results are shown for three examinees. Depending on their location on the General factor, as it was estimated by the General factor CAT, each examinee potentially received both a different number of items and a different subset of items. Figure 1a shows that Examinee 1 received 15 items, Examinee 2 received 16 items, and Examinee 3 answered 13 items. Of the items answered by each examinee, different subsets of items were associated with Content Factor 1 in the bifactor structure—these item numbers are shown in blue. Examinee 1 answered 5 items from Content Scale 1, Examinee 2 answered 7 items (with 1 item common with Examinee 1), and Examinee 3 answered only 3 items from Scale 1. Figure 1b shows the five Scale 1 items answered by Examinee 1. These items were used, in conjunction with their IRT discrimination parameters on Scale 1, to derive an initial θ estimate for Examinee 1 on Scale 1, which was used to select the first item in Scale 1; the Scale 1 CAT than proceeded from that starting θ estimate.

Figure 1
Illustration of Item Usage in the Bifactor CAT Algorithm

a. Items Administered by a CAT on the General Factor (G) to Three Hypothetical Examinees



b. Using Scale 1 Items From the G Factor to Begin the Scale 1 CAT for Examinee 1



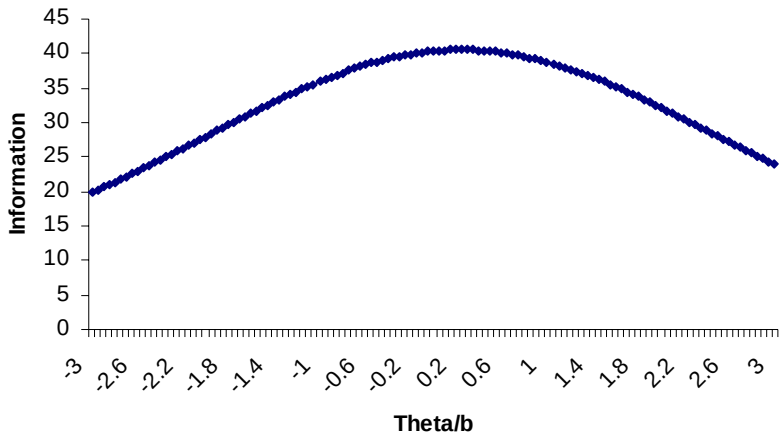
Method

Data

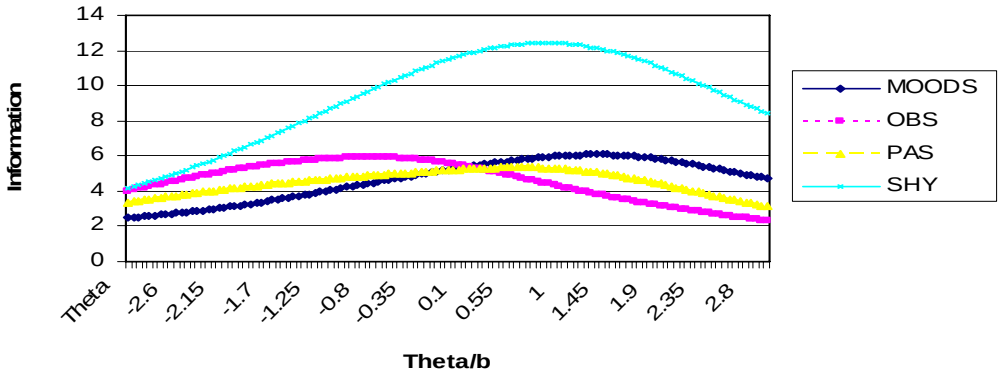
Instrument. The Mood-Anxiety Spectrum Scales (MASS: Cassano, Michelini, Shear, et al. 1997; Frank, Cassano, Shear, et al., 1998) provided the data for this research. The MASS consists of 626 items that score on a General Scale and four content scales. All items are scored dichotomously, with higher scores reflecting higher levels of pathology. Of the 626 items, 615 were used in this study (the remainder of the items had insufficient variance to be useful). All 615 items were scored for the General scale. The content scales were Mood (155 items), Obsessive-Compulsive (OBS: 183 items), Panic-Agoraphobia (PAS: 113 items), and Social Phobia (Shy: 164 items). Figure 2 shows the scale information functions for each of the five MASS scales based on the bifactor analysis results.

Figure 2
Scale Information Functions for Five MASS Scales

a. General Factor Scale



b. Content Scales



Participants. Item response data were obtained from four groups of examinees, primarily as part of the Mental Health Computerized Adaptive Testing (MHCAT) study (Gibbons et.al., 2008):

1. *Item calibration group.* Data were from 800 participants in outpatient treatment for a mood or anxiety disorder at the Western Psychiatric Institute and Clinic in Pittsburgh, PA The MASS was computer-administered to this group based on an optimal balanced incomplete block design that was designed to maximize the number of pairings of 616 of the 626 items while minimizing the number of items administered to each subject. The FastTEST Professional Testing System (Weiss, 2006) was used to create 36 different test forms, each consisting of 154 items extracted from the four MASS subscales. Items were administered to 36 randomly assigned subgroups of research participants (sample sizes for the 36 groups varied from 17 to 28 with a median of 22).
2. *Post-hoc CAT calibration group (PH-1*). Data for this group included complete responses obtained from paper-and-pencil administration of all 626 MASS items from 148 depressed patients in an earlier study conducted jointly by the Universities of Pittsburgh (N = 90) and Pisa, Italy (N= 58). These item responses were used as the basis for initial post-hoc simulations of the bifactor CAT algorithm. The objective of these simulations was to use the bifactor CAT algorithm to determine the numbers of items required to obtain CAT θ estimates that correlated .90 or above with full-scale θ estimates for the five MASS scales. A number of standard error (SEM) termination criteria were examined until appropriate SEMs were obtained that satisfied the correlation criterion for all five scales.
3. *Post-hoc cross-validation group (PH-2).* Data for this group were obtained from complete responses to the MASS from computer administration of 615 MASS items. The 204 participants were from data collected by MHCAT. This group was used to cross-validate the post-hoc simulations in the PH-1 group. In this group the bifactor CAT algorithm was implemented with the SEM termination values identified in the PH-1 group and correlations between the CAT and full-scale, $\hat{\theta}s$ as well as test length for each scale, were observed.
4. *Live-testing bifactor CAT group.* Participants in this group were a subset of 156 examinees from the PH-2 group who returned for an additional testing session an average

of 5.5 months after completing the full MASS at their earlier session. For this group of 156 examinees, the bifactor CAT was administered live using the FastTEST Professional Testing System (Weiss, 2006). Data from the first PH-2 testing session were retrieved for these 156 examinees to provide test-retest results based on post-hoc bifactor CAT simulation at the first session and live bifactor CAT administration at the second session; the first session data for this reduced set of examinees are referred to as the PH-2-R group.

Procedure

Post-hoc simulations were based on complete sets of item responses to the 615 MASS items (one item was eliminated because it had no variance). Using the bifactor CAT algorithm described above, the program POSTSIM3 (Weiss, 2008) used the existing item responses in the CAT algorithm to "administer" a CAT to each examinee. The resulting CAT $\hat{\theta}s$ were compared to the full-scale $\hat{\theta}s$ for each of the five MASS scales, and these CAT $\hat{\theta}s$ were compared to the full-scale $\hat{\theta}s$ estimated from all the item responses in each scale.

Live CAT administration was also based on the bifactor CAT algorithm, but MASS items were stored in the testing computer and items were presented to examinees one at a time. Item responses were obtained by mouse clicks on the "yes"/"no" item responses on the screen, which highlighted when clicked. Examinees were allowed to change item responses until they clicked on a green arrow in the top corner of the screen. The response was then scored, θ was estimated, and the next item was selected and presented. Inter-item delays were always a second or less.

Data Analysis

Of interest was the efficiency of the bifactor CAT algorithm in recovering full-scale θ estimates with minimal loss of measurement quality for each of the five MASS scales separately and for all scales combined. The following evaluative criteria were computed for each scale:

1. Pearson correlations between bifactor CAT $\hat{\theta}s$ $\left(\hat{\theta}_{\mathrm{C}}\right)$ and full-scale $\hat{\theta}s$ $\left(\hat{\theta}_{\mathrm{F}}\right)$.

2. Bias: The mean signed difference between CAT and full-scale θ estimates,

$$\text{bias} = \frac{\sum_{i=1}^{N}\left(\hat{\theta}_{i\mathrm{C}} - \hat{\theta}_{i\mathrm{F}}\right)}{N} \tag{3}$$

3. Accuracy: The mean absolute difference between CAT and full-scale θ estimates.

$$\text{accuracy} = \frac{\sum_{i=1}^{N}\left|\hat{\theta}_{i\mathrm{C}} - \hat{\theta}_{i\mathrm{F}}\right|}{N} \tag{4}$$

4. The average number of items required by CAT to recover full-scale θ estimates with a correlation of .90 or higher, or with a pre-specified SEM.

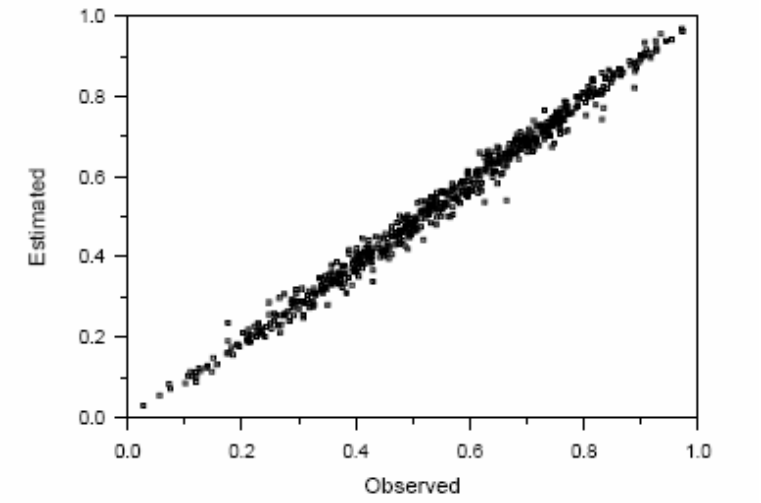
5. Relationship between live-testing and post-hoc simulation CAT results.

Results

Fit of the Bifactor Model

Both unidimensional and bifactor models (using the content scales as secondary dimensions) were fitted to the 616 item responses obtained from the 800 research participants in the item calibration group. The bifactor model provided acceptable model-data fit and significantly improved model fit over the unidimensional model (χ^2 = 2,955, df = 616, p < 0.001), supporting the scale's multidimensional structure. That is, the bifactor results supported the contribution of each of the four MASS domains in addition to the primary domain for accounting for the MASS's underlying factor structure. Figure 3 displays the relationship between the observed proportion endorsed and the proportion endorsed estimated by the bifactor model for the 616 items, reflecting the excellent fit of the bifactor model to the item responses.

Figure 3
Relationship Between Observed Proportion Endorsed and Estimated Proportion Endorsed From the Bifactor Model for 616 MASS Items



Performance of the Bifactor CAT

General scale. Results for all scales are shown in Table 2. For the General scale in the PH-1 group, the post-hoc simulation analyses revealed that on average, 24.4 items (range of 18 – 55 items) were required to achieve $r\left(\hat{\theta}_C, \hat{\theta}_F\right)$ = .922. The resulting target SEM was approximately .30. When this SEM was used in the PH-2 cross-validation group, $r\left(\hat{\theta}_C, \hat{\theta}_F\right)$ = .933 was obtained with an average of 23.76 items (range 18 – 77). Bias was low for both groups: –.277 for PH-1 and –.095 for PH-2. Accuracies of .375 and .371, respectively, were consistent with the SEM termination value of .30. Test length mean percentage reductions for both groups were 96%. When the SEM termination criterion for the General scale was applied in live testing, an average of 30.64 items were administered, with a somewhat larger range than for both post-hoc simulations. However, the mean percent reduction in test length was 95% in live testing versus 96% for the same examinees (PH-2-R) under post-hoc simulation.

Table 2
Correlation of $\hat{\theta}_C$ With $\hat{\theta}_F$, Bias and Accuracy, Mean and Range of Number of Items Administered, and Percent Mean Reduction in Number of Items for Five Scales

Scale and Group	$r(\hat{\theta}_C, \hat{\theta}_F)$	Bias	Accuracy	Number of Items		
				Mean	Range	Reduction
General Scale (615 Items, SEM = .30)						
PH-1	.922	–.277	.375	24.45	18 – 55	96.02%
PH-2	.933	–.095	.371	23.76	18 – 77	96.14%
PH-2-R	.931	–.063	.371	23.86	18 – 77	96.12%
Live CAT	--	--	--	30.64	18 – 184	95.02%
Scale 1: Mood (155 Items, SEM = .35)						
PH-1	.924	–.137	.334	27.16	16 – 60	82.48%
PH-2	.972	.098	.291	48.61	16 – 155	68.64%
PH-2-R	.973	.078	.282	50.06	16 – 155	67.60%
Live CAT		--	--	27.26	14 – 62	82.41%
Scale 2: OBS (183 Items, SEM = .475)						
PH-1	.931	–.219	.453	29.70	9 – 67	83.77%
PH-2	.915	–.048	.484	18.99	8 – 67	89.62%
PH-2-R	.902	–.034	.508	18.53	8 – 67	89.87%
Live CAT	--	--	--	20.97	9 – 84	88.55%
Scale 3: PAS (113 Items, SEM = .40)						
PH-1	.948	–.270	.377	23.40	10 – 79	79.61%
PH-2	.958	.025	.341	21.94	10 –79	80.58%
PH-2-R	.956	.045	.343	22.03	10 –79	80.50%
Live CAT	--	--	--	18.62	6 – 77	83.52%
Scale 4: Shy (164 Items, SEM = .35)						
PH-1	.953	–.215	.412	17.63	11 – 61	89.25%
PH-2	.968	–.119	.351	28.28	12 – 164	82.75%
PH-2-R	.969	–.094	.341	27.52	12 – 164	83.20%
Live CAT	--	--	--	34.38	19 – 136	79.04%

Content scales. Results for the four content scales were similar to those obtained for the General scale. Correlations between $\hat{\theta}_C$ and $\hat{\theta}_F$ were all in the .90s, ranging from .902 to .973. Bias for all scales and across all groups was negligible, with a slight predominance of small negative values. Accuracies were generally consistent with the SEM termination values used for each scale. Test length mean percentage reductions were somewhat smaller than for the General scale, ranging from about 68% to about 90%. The mean number of items administered in the live CAT was generally consistent with the mean number administered for the PH-2-R group; however, for Scale 2 the mean number of items in post-hoc simulation

was 50.06 whereas for the same examinees in live testing a mean of only 27.26 items was required to terminate the bifactor CAT on that scale.

All scales combined. Table 3 shows measures of central tendency for number of items administered to examinees by the bifactor CAT algorithm across all scales combined. With the exception of the mean for live CAT administration, mean, median, and modal percent reductions were 80% and above for all groups. Median number of items required to measure examinees on all scales was 100 for live CAT and 93.5 and 94.5 for the two post-hoc groups. Modes and means of number of items were slightly higher in live testing than in post-hoc simulation, with about a 6% maximum difference for the means.

Table 3
Mean, Median, and Mode of Total Number of Items Administered by CATs (*n*) on the Five Scales and Percent Reduction From 615 Items, for Three Groups

	Mean		Median		Mode	
Group	*n*	%	*n*	%	*n*	%
PH-1	97.90	84.1%	93.5	84.8%	77	87.5%
PH-2	118.1	80.1%	94.5	84.6%	73	88.1%
Live CAT	131.8	78.6%	100	82.1%	92	85.0%

Actual test administration times were recorded for the 156 examinees in group PH-2-R who took the 626 MASS items by computer and who later took the bifactor CAT. On initial test administration, the mean testing time was 114.81 minutes, with a SD of 46.13 minutes. When the examinees later completed the bifactor CAT, mean administration time was 22.05 minutes (SD = 11.66). This was an average time reduction of 82%, or a mean time saving of 92.76 minutes.

Correlations of θ estimates from post-hoc simulation and live CAT. Table 4 shows the correlations among $\hat{\theta}$s for the PH-2-R group from both post-hoc and live CAT administration. The first row shows the test-retest correlations between the two CAT $\hat{\theta}$s over the average five-month period. These correlations ranged from .749 to .827, which compare favorably with the test-retest correlations between the live CAT $\hat{\theta}$ and the full post-hoc $\hat{\theta}$, shown in the second row. The former correlations were no more than .05 different from the latter, with the largest difference observed for the General scale. The last row of Table 4 shows the correlations between $\hat{\theta}$s for this group when there was no time interval between the $\hat{\theta}$s .

Table 4
Correlations Among CAT $\hat{\theta}$s and Full Scale $\hat{\theta}$s From the Five Scales for the PH-2-R Group Based on Post-Hoc CAT and Live CATs (N = 156)

	Scale				
θ Estimates	General	Mood	OBS	PAS	Shy
Live CAT $\hat{\theta}$ with post-hoc CAT $\hat{\theta}$	.776	.826	.797	.749	.827
Live CAT $\hat{\theta}$ with full post-hoc $\hat{\theta}$	.826	.832	.829	.767	.847
Post-hoc CAT $\hat{\theta}$ with full post-hoc $\hat{\theta}$	.931	.973	.902	.956	.967

Conclusions

The bifactor model provided a good fit to the MASS scale items. The fit of the bifactor model was significantly better than the fit of a unidimensional model, and the estimated proportions of keyed responses from the bifactor model closely fit the observed proportions.

The CAT bifactor algorithm resulted in very substantial reductions in numbers of items (80% to 95%) in both post-hoc simulation and live bifactor CATs. while producing CAT θ estimates that correlated above .90 for all five MASS scales with θ estimates from the full sets of scale items. Comparisons of the results from the post-hoc simulation groups indicated that standard error termination criteria identified in the first group that resulted in correlations above .90 cross-validated well in the second post-hoc group, both in terms of correlations and reductions in numbers of items.

The results of the post-hoc simulation generally well predicted the results of live bifactor CAT administration, thus supporting the usefulness of post-hoc simulation in the process of developing and implementing operational CATs. Results from the live testing, based on a test-retest design, showed high correlations between CAT θ estimates from the post-hoc simulation and retests of live CATs an average of about five months later Across all scales of the MASS, number of items need to obtain θ estimates was reduced by the bifactor CAT algorithm about 80% to 85%, with a mean testing time reduction of 82%, which translated to a mean saving of examinee testing time of approximately 93 minutes over administration of the full set of MASS items.

References

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

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

Cassano, G. B., Michelini, S., Shear, M. K., et al. (1997). The panic-agoraphobic spectrum: A descriptive approach to the assessment and treatment of subtle symptoms. *American Journal of Psychiatry*, 154 (suppl 6), 27-38.

Frank E., Cassano, G. B., Shear, M.K., et al. (1998). The spectrum model: A more coherent approach to the complexity of psychiatric symptomatology. *CNS Spectrums*, *3*, 23-34.

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

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31,* 4–19.

Gibbons R.D., Weiss D.J., Kupfer D.J., Frank E., Fagiolini A., Grochocinski V.J., Bhaumik D.K., Stover A. Bock R.D., Immekus J.C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59,* 361-368.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M. D., & McKinley, R. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Segall, D. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331-354.

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53-73). Norwell MA: Kluwer.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics, 24,* 398-412.

Weiss, D. J. (2006). *Manual for the FastTEST Professional Testing System, Version 2.* St. Paul MN: Assessment Systems Corporation.

Weiss, D. J. (2008). *Manual for POSTSIM 3: Post-hoc simulation of computerized adaptive testing. Version 3.0.* St. Paul MN: Assessment Systems Corporation.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (2002). *TESTFACT* [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.