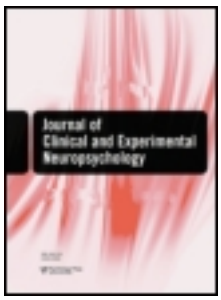

Journal of Clinical and Experimental Neuropsychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ncen20

Parallel psychometric and cognitive modeling analyses of the Penn Face Memory Test in the Army Study to Assess Risk and Resilience in Servicemembers

Michael L. Thomas [a], Gregory G. Brown [a b], Ruben C. Gur [c d], John A. Hansen [c d], Matthew K. Nock [e], Steven Heeringa [f], Robert J. Ursano [g] & Murray B. Stein [a]

[a] Department of Psychiatry, University of California, San Diego, San Diego, CA, USA

[b] VA San Diego Healthcare System, San Diego, CA, USA

[c] Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA

[d] Philadelphia VA Medical Center, Philadelphia, PA, USA

[e] Department of Psychology, Harvard University, Cambridge, MA, USA

[f] Institute for Social Research, University of Michigan, Ann Arbor, MI, USA

[g] Department of Psychiatry, Uniformed Services University of the Health Sciences, Bethesda, MD, USA

Version of record first published: 05 Feb 2013.

To cite this article: Michael L. Thomas , Gregory G. Brown , Ruben C. Gur , John A. Hansen , Matthew K. Nock , Steven Heeringa , Robert J. Ursano & Murray B. Stein (2013): Parallel psychometric and cognitive modeling analyses of the Penn Face Memory Test in the Army Study to Assess Risk and Resilience in Servicemembers, Journal of Clinical and Experimental Neuropsychology, DOI:10.1080/13803395.2012.762974

To link to this article: http://dx.doi.org/10.1080/13803395.2012.762974

JOURNAL OF CLINICAL AND EXPERIMENTAL NEUROPSYCHOLOGY, 2013
http://dx.doi.org/10.1080/13803395.2012.762974

Parallel psychometric and cognitive modeling analyses of the Penn Face Memory Test in the Army Study to Assess Risk and Resilience in Servicemembers

Michael L. Thomas[1], Gregory G. Brown[1,2], Ruben C. Gur[3,4], John A. Hansen[3,4], Matthew K. Nock[5], Steven Heeringa[6], Robert J. Ursano[7], and Murray B. Stein[1]

[1]Department of Psychiatry, University of California, San Diego, San Diego, CA, USA
[2]VA San Diego Healthcare System, San Diego, CA, USA
[3]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA
[4]Philadelphia VA Medical Center, Philadelphia, PA, USA
[5]Department of Psychology, Harvard University, Cambridge, MA, USA
[6]Institute for Social Research, University of Michigan, Ann Arbor, MI, USA
[7]Department of Psychiatry, Uniformed Services University of the Health Sciences, Bethesda, MD, USA

Objective: The psychometric properties of the Penn Face Memory Test (PFMT) were investigated in a large sample (4,236 participants) of U.S. Army Soldiers undergoing computerized neurocognitive testing. Data were drawn from the initial phase of the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS), a large-scale study directed towards identifying risk and resilience factors for suicidal behavior and other stress-related disorders in Army Soldiers. In this paper, we report parallel psychometric and cognitive modeling analyses of the PFMT to determine whether ability estimates derived from the measure are precise and valid indicators of memory in the Army STARRS sample. *Method:* Single-sample cross-validation methodology combined with exploratory factor and multidimensional item response theory techniques were used to explore the latent structure of the PFMT. To help resolve rotational indeterminacy of the exploratory solution, latent constructs were aligned with parameter estimates derived from an unequal-variance signal detection model. *Results:* Analyses suggest that the PFMT measures two distinct latent constructs, one associated with memory strength and one associated with response bias, and that test scores are generally precise indicators of ability for the majority of Army STARRS participants. *Conclusions:* These findings support the use of the PFMT as a measure of major constructs related to recognition memory and have implications for further cognitive–psychometric model development.

Keywords: Psychometric modeling; Item response theory; Cognitive modeling; Penn Face Memory Test; Army Study to Assess Risk and Resilience in Servicemembers.

Psychometric analyses provide the bedrock for accurate test interpretation in clinical neuropsychological practice and research (Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Loevinger, 1957). It is essential that studies of mental health provide clear evidence that test scores are reliable and valid indicators of psychological constructs. Modern psychometric techniques including multidimensional item response theory (MIRT; see Reckase, 2009) constitute the preeminent methodology for such analyses. In this paper, we explore MIRT models of the Penn Face Memory Test (PFMT; Gur et al., 1993; Gur, Jaggi, Shtasel, & Ragland, 1994; Gur et al., 2001; Gur et al., 1997;

Army STARRS was sponsored by the Department of the Army and funded under cooperative agreement U01MH087981 with the U.S. Department of Health and Human Services, National Institutes of Health, National Institute of Mental Health (NIH/NIMH). The contents are solely the responsibility of the authors and do not necessarily represent the views of the Department of Health and Human Services, NIMH, the Department of the Army, or the Department of Defense.

Address correspondence to Gregory G. Brown, San Diego VA Medical Center, Psychology Service (116b), 3350 La Jolla Village Drive, San Diego, 92161-0002, CA, USA (E-mail: gbrown@ucsd.edu).

Gure et al., 2010) and improve interpretations of parameter estimates using results from a parallel cognitive modeling analysis. These results are used to assess the measurement precision of the PFMT as a component of the neuropsychological battery being administered to U.S. Soldiers in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS; Army STARRS, 2011). Army STARRS is a large-scale study directed towards the identification of biological and psychosocial risk and resilience factors for suicidal behavior and other stress-related disorders. The present study serves to establish the measurement properties of the PFMT and provides a basis for future development of predictive models.

Context of analyses

Army STARRS is a collaborative research effort in response to the alarming increase in suicide rates among Army Soldiers over the past several years (U.S. Army, 2010). The study, which is divided into several components, includes historical data, prospective data collection, and biological sample collection across a number of substudies. The New Soldier Study—a component of Army STARRS—involves the administration of computerized questionnaires, including psychiatric symptom inventories, psychological assessments, and neuropsychological tests to Army Soldiers at the onset of basic training. The final sample for the New Soldier Study is expected to comprise tens of thousands of participants. Data collection for the project began in 2011 and was still underway at the time of writing this manuscript.

The neuropsychological battery being administrated to Soldiers establishes baseline cognitive functioning with the future aim of identifying markers of risk and resilience. Tests were designed to provide efficient assessment of neurocognitive domains that can be linked to specific brain systems (see Penn Computerized Neurocognitive Battery; Gur, Erwin, & Gur, 1992; Gur et al., 2012; Gur et al., 2010). Army STARRS test data will be linked with demographic data, measures of mental health, and genetic markers collected in the final sample. Therefore, it is important that these data are reliable and valid indicators of Army Soldiers' neurocognitive functioning.

Model development

Analyses of data drawn from psychological instruments rely on models that relate test items to latent abilities. Reliability coefficients from both classical test theory (e.g., Cronbach's alpha) and common factor theory (e.g., omega) are based on specific assumptions about underlying measurement structures (e.g., tau-equivalent or common factor; see McDonald, 1999). Verification of such assumptions, including a proper understanding of the test space, pattern of loadings, and factor correlations, is needed to derive accurate reliability and validity coefficients. Unfortunately, the process of mapping relations between observed data and latent abilities in psychological assessment is complex. There is no guarantee, for example, that item success within a specific scale requires just a single cognitive ability; there is also no guarantee that success is determined equally by the same abilities for all test takers (see Haynes, Smith, & Hunsley, 2011). In neuropsychological assessment, this opacity is worsened when separate cognitive processes are involved in task performance (see Brandt, 2007), further hindering the goal of establishing the construct validity of test scores (see Strauss & Smith, 2009).

A combination of theory and empirical evidence is needed to overcome the difficulties associated with measuring latent abilities. Combining principles from correlational and experimental psychology in model development (cf. Cronbach, 1957) is exemplified in current methodology known as cognitive psychometrics (see Batchelder, 2010). In analyses of the PFMT, cognitive–psychometric modeling can be used to explain unknown aspects of the test's latent factor structure with experimentally derived cognitive architecture. To do so, we start with a firm understanding of the PFMT's theoretical bases.

The PFMT was designed to measure visual episodic memory for unstructured stimuli, a latent ability sensitive to right temporal lobe dysfunction (Gur et al., 1993; Gur et al., 1994). The task parallels aspects of commonly administrated measures of verbal episodic memory (e.g., Delis, Kramer, Kaplan, & Ober, 2000), requiring examinees to first memorize a list of faces and then to identify test items as either targets (old faces) or distractors (new faces). Although there is strong evidence that neuropsychological tests can dissociate visual episodic memory from other cognitive abilities (Cabeza & Kingstone, 2006; Carroll, 1993; Lezak, Howieson, Loring, Hannay, & Fischer, 2004), it is more difficult to delineate specific subsystems and/or levels of processing within the domain itself (e.g., Baddeley & Hitch, 1974; Cowan, 1988). Indeed, it is well known that episodic memory tasks activate several regions of the prefrontal cortex and medial temporal lobe in brain imaging studies

(Cabeza & Kingstone, 2006; Gur et al., 1997; Kelley et al., 1998).

A parsimonious explanation of recognition memory test performance is offered by cognitive modeling studies based on signal detection theory (see Green & Swets, 1966; Wickens, 2002), which have suggested that differences between target and distractor responses can be attributed to two general constructs: (a) response bias; and (b) memory strength. Interplay between the constructs can be modeled using two unequal-variance Gaussian distributions (one for targets and one for distractors) and a set of decision criteria (Wixted, 2007). The free parameters of the model are d', distance between the means of the target and distractor distributions; ${\sigma^2}_{\mathrm{T}}$, variance of the target distribution; and C, response bias (the distance between criterion placement and the midpoint of d'; see Snodgrass and Corwin, 1988; additional details provided in the Appendix). Memory strength is related to d' and ${\sigma^2}_{\mathrm{T}}$, where larger values of d' and smaller values of ${\sigma^2}_{\mathrm{T}}$ equate with better accuracy. Response bias is represented by the C parameter, where higher values of C lead to conservative responding (favoring distractors), and lower values of C lead to liberal responding (favoring targets). Note that this definition of response bias is not synonymous with poor effort. Because the signal detection model has been widely discussed in the literature, specific formulas are not presented here. Importantly, parameters of the model can be conceptualized at the level of individual respondents (e.g., Ingham, 1970) so that differences in d', ${\sigma^2}_{\mathrm{T}}$, and C can be directly related to parameter values from psychometric models to better understand a test's internal structure.

The psychometric models used in our analyses are all based on the multidimensional extension of an item response theory graded response model (M-GRM; de Ayala, 1994; Muraki & Carlson, 1995; Samejima, 1997). The M-GRM is a psychometric framework for test data comprising ordered categorical responses. For the PFMT, examinees must choose from four responses for both target and distractor items indicating their confidence in seeing the test item previously: "definitely yes," "probably yes," "probably no," or "definitely no." The M-GRM first relies on a logistic distribution to model the probability that examinee i (N total) will respond with the kth or higher category (K total) to item j (J total) as a function of the item category threshold (d_{jk}), item discrimination (a_{jm}), and examinee ability[1] (θ_{im}) parameters by

$$P(u_{ij} \geq k | d_{jk}, \mathbf{a}_j, \boldsymbol{\theta}_i) = \frac{\exp\left(\sum_{m=1}^{M} a_{jm}\theta_{im} + d_{jk}\right)}{1 + \exp\left(\sum_{m=1}^{M} a_{jm}\theta_{im} + d_{jk}\right)}, \quad (1)$$

where $\mathbf{a}_j$ and $\boldsymbol{\theta}_i$ are vectors with M elements (one per factor), and u_{ij} is the realization of the response variate. From these values, the probabilities of responding in each of the categorical response options are then given by

$$P(u_{ij} = 1 \mid d_{j1}, \mathbf{a}_j, \boldsymbol{\theta}_i) = 1 - P(u_{ij} \geq 2 \mid \boldsymbol{\theta}_i, \mathbf{a}_j, d_{j1}) \quad (2)$$

$$\ldots$$

$$P(u_{ij} = k \mid d_{jk}, \mathbf{a}_j, \boldsymbol{\theta}_i) = P(u_{ij} \geq k \mid \boldsymbol{\theta}_\iota, \mathbf{a}_j, d_{jk}) - P(u_{ij} \geq k+1 \mid \boldsymbol{\theta}_\iota, \mathbf{a}_j, d_{jk+1}) \quad (3)$$

$$\ldots$$

$$P(u_{ij} = K \mid d_{jK}, \mathbf{a}_j, \boldsymbol{\theta}_i) = P(u_{ij} \geq K | \boldsymbol{\theta}_\iota, \mathbf{a}_j, d_{jK}). \quad (4)$$

As can be seen, the probability of responding with the first categorical response option is simply one minus the probability of responding with any other option. The probability of responding with an intermediate categorical response option is the difference of being in k or higher and $k + 1$ or higher for each category. Finally, the probability of responding with the last categorical response option is just the probability of responding with this option alone. Equations 2, 3, and 4 rely on the restriction that the item threshold parameters (d_{jk}) are strictly ordered from largest to smallest from the first to last categorical response options, respectively. This assures that the response probabilities for a fixed ability level decrease from one category to the next (i.e., that correct PFMT response options are more difficult than incorrect PFMT response options).

The item threshold (d_{jk}) parameter is inversely related to item difficulty (i.e., higher values indicate lower difficulty). The item discrimination (a_j) parameter can be interpreted similarly to a factor loading or an item-by-total biserial correlation. In a graphical representation of the item response theory (IRT) model, the d_{jk} and a_j parameters determine the locations and slopes of the items' category response functions. An example is illustrated in the top panel of Figure 1 for a hypothetical PFMT item. The plot shows that the probability of responding in the kth of K categories (y-axis) is a function of ability (x-axis). The exact shapes of the curves depend on the item parameters. The θ

[1]The terms "ability," "factor," and "construct" are similarly used to indicate a latent variable.

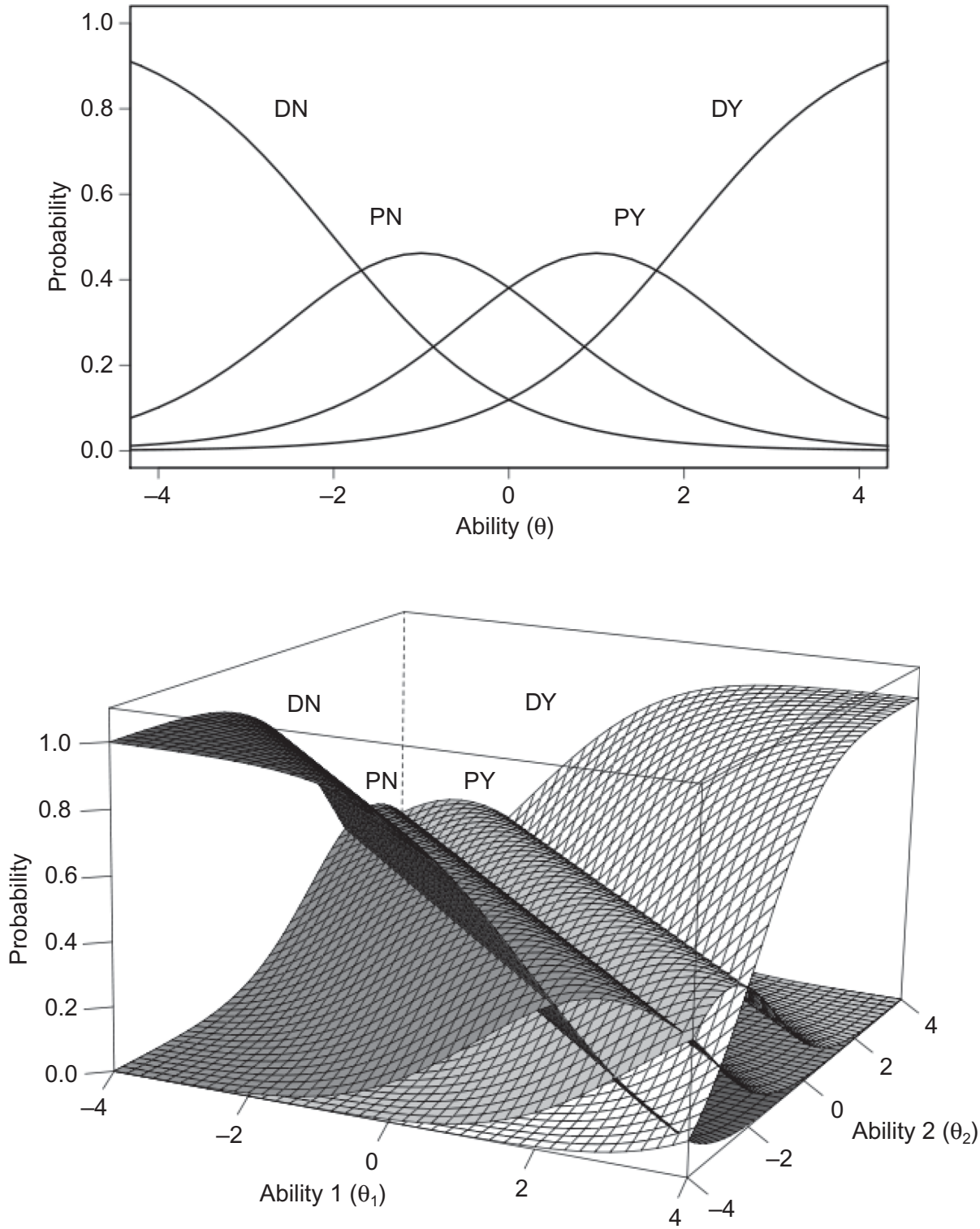

Figure 1. Category response curves for hypothetical unidimensional (top) and multidimensional (bottom) Penn Face Memory Test items. DY = "definitely yes," PY = "probably yes," PN = "probably no," or DN = "definitely no."

parameter represents examinees' standings on the cognitive ability assumed to influence individual differences in item responses. The model represents item response variance as a function of examinee ability, item threshold (inverse of difficulty), and item discrimination.

The vector notations $\mathbf{a}_j$ and $\boldsymbol{\theta}_i$ used in Equations 1, 2, 3, and 4 imply that multiple

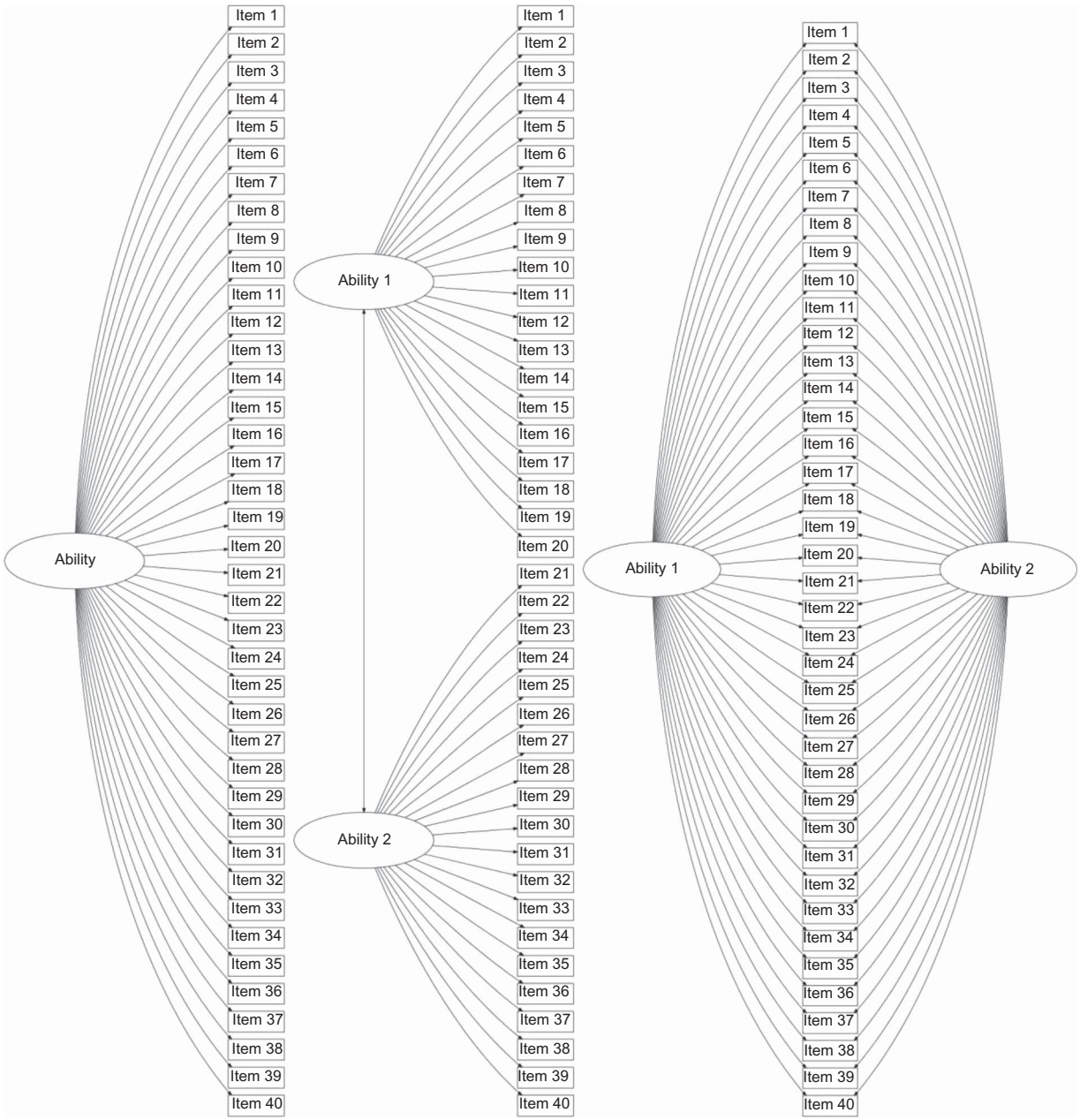

Figure 2. Unidimensional (left), between-item multidimensional (center), and within-item multidimensional (right) models.

cognitive abilities can influence responses. In the case where all items can discriminate between distinct levels of all abilities, the model is said to have within-item multidimensional structure. When items can discriminate just one of two or more abilities, the model is said to have between-item multidimensional structure. And, when all items can discriminate just a single ability, the model is said to have unidimensional structure. The within-item multidimensional model is the most general, and the unidimensional model is the most restrictive. These three nested variants are depicted in Figure 2. Examples of within-item multidimensional item category response functions are provided in the bottom panel of Figure 1. Note that the hypothetical item can discriminate between levels of multiple abilities (i.e., the curvatures of the category response functions change with level of ability along both axes). Several sources can be consulted for further description of item response theory models in psychological and clinical assessment (Embretson & Reise, 2000; Reise & Waller, 2009; Thomas, 2011).

Measurement precision

Item response theory (IRT) offers a robust and comprehensive alternative to classical test theory (CTT). Both IRT and CTT provide descriptions of and formulas for measurement precision; however, there are important differences among them. In CTT, standard error (SE_{tot}) is derived from test reliability by

$$SE_{tot} = \sigma_{tot}\sqrt{1-\rho}, \tag{5}$$

where σ_{tot} is the population standard deviation of the total number correct true score, and ρ is the population reliability coefficient (Magnusson, 1966). Standard error in IRT is derived from item information—the instantaneous change in the probability of a correct response on a particular item at a particular ability level normalized by item variance at the ability (Lord, 1980). An item where a small change from a particular ability value produces a large normalized change in the probability of a correct response is informative at that ability. This occurs when item difficulty is closely matched with ability and when item discrimination is high (Baker & Kim, 2004). Test information at a particular ability (TI_θ) is simply the sum of item information. From this, standard error as a function of ability, SE_θ, is given by

$$SE_\theta = \frac{1}{\sqrt{TI_\theta}}. \tag{6}$$

Whereas standard expositions of CTT provide a single estimate of total score measurement error based on reliability, IRT provides distinct estimates of measurement error that can vary depending on ability; that is, potentially unique values of SE_θ can be assumed for unique values of θ.

It has been shown that CTT could be extended to include the concept of test information and that IRT could be extended to include the concept of reliability (e.g., Mellenbergh, 1996). For example, reliability could be calculated in IRT across a range of ability estimates using the average standard error (Andrich, 1988). Yet, in practice, information and reliability are rarely both computed from a particular test model. CTT estimates of measurement precision are almost always reported as true score reliability, and IRT estimates of measurement precision are almost always reported as information. Furthermore, because CTT estimates of reliability are for particular items administered to particular groups of examinees, standard error is often reported as a single value that confounds examinee and test properties (Embretson & Reise, 2000). Within the IRT framework, item properties, such as item difficulty, are modeled in addition to ability, facilitating estimates of a test's psychometric properties that are independent from examinee properties (Embretson & Reise, 2000).

As a practical tool for researchers, IRT could easily offer both SE_θ and ρ for evaluating measurement precision. Evaluations of SE_θ involve comparisons among differing levels of ability. Usually, graphical representations of the SE_θ function are used to determine ability locations with relatively better or worse measurement precision. By doing so, a test evaluator can determine whether a test is optimized for measurement within a specific population and a specific range of ability. Cutoffs for acceptable SE_θ, however, are less commonly offered, which can sometimes lead to ambiguous interpretations of data. Unlike SE_θ, acceptable cutoffs for ρ are commonly offered (.60 to .80 depending on the specific type of coefficient; see Haynes et al., 2011). Typically, we should expect that a favorable ρ should accompany a favorable SE_θ function with respect to a specific population. That is, reliability should be high when areas of low measurement error overlap with areas of high density in ability. However, because reliability is a reflection of average standard error, implying that it is possible to find equivalent reliabilities for tests with very different standard error functions, both indices of measurement precision should be evaluated.

Study goals

The first goal of this study was to find an appropriate measurement model for the PFMT. This was accomplished by fitting nested measurement structures based on the M-GRM to a training data set and then cross-validating the best fitting model in a validation data set. To improve interpretability, the psychometric results were then compared to cognitive modeling parameters derived from the unequal-variance signal detection model. Our second goal was to evaluate the measurement precision of PFMT latent ability scores. To do so, we compared SE_θ values and evaluated ρ over the range of ability observed in Army STARRS.

METHOD

Participants

Army Soldiers were recruited to volunteer without compensation for the Army STARRS New Soldier Study at the start of basic training. The current

sample comprises participants from three Army bases in the United States tested between February and June of 2011. All Soldiers were asked to provide informed, written consent prior to participation in research. Army commanders provided sufficient time to complete all surveys and tests, which were administered in a group format using laptop computers. Research proctors monitored the testing environment and assisted with questions and technical difficulties. Surveys and tests were administered in a fixed order in 90-min sessions over two days of testing. The PFMT was administered on the second day of testing.

A total of 5,551 Army Soldiers recruited into the study participated in the New Soldier Survey during the specified time frame. Of these, approximately 15% did not complete the PFMT (e.g., some did not show for the second day of testing), and another 8% were removed due to technical problems or potentially invalid data. Specifically, during one administration a subset of participants experienced a testing software malfunction. The software anomalies were corrected, and incomplete response profiles were flagged as errant. Also, Soldiers who provided the same response to 12 or more consecutive items were flagged as errant. This automated rule for identifying invalid profiles on the PFMT was developed based on previous large-scale studies using the Penn Computerized Neurocognitive Battery (Gur et al., 2012). Profiles flagged as errant were not used for psychometric modeling. Automated quality control is required due the complexities of validating thousands of individual response profiles in large-scale studies. Other large-scale studies have reported similar exclusion rates due to technical problems (Hoerger, Quirk, & Weed, 2011).

These procedures resulted in a final sample of 4,236 participants for psychometric modeling. Demographic characteristics of participants with PFMT data and participants without PFMT data are shown in Table 1. No group differences were found for age and handedness; however, there tended to be fewer males, better educational attainment, and less ethnic and racial diversity in the group of participants with valid PFMT data.

TABLE 1
Demographic characteristics of participants with Penn Face Memory Test data present versus absent or excluded

Characteristic	*Data present*	*Data absent/excluded*	p^a
Age			
M	21.49	21.35	.30
SD	4.34	4.44	
Gender			
Male	76	84	<.001
Education			
GED	7	9	.04
HS Diploma	46	54	<.001
Post high school no certificate	26	18	<.001
Post high school certificate	4	5	.41
2-year degree	7	8	.14
4-year degree	9	5	<.001
Some graduate	2	2	.91
Handedness			
Right-handed	86	87	.47
Left-handed	11	10	.48
Either	3	3	.94
Race			
White	68	62	<.001
Black	21	23	.04
American Indian or Alaskan Native	2	2	.90
Asian	3	3	.62
Native Hawaiian or Pacific Islander	1	1	.25
Other	6	9	<.001
Ethnicity			
Not Spanish/Hispanic/Latino	86	82	.003
Mexican	5	7	.005
Puerto Rican	4	5	.17
Cuban	<1	1	.51
Other Spanish/Hispanic/Latino	5	5	.86

[a]Significance (p) based on t tests for continuous variables and χ^2 tests for categorical variables.

Measure

The PFMT was administered as part of a neurocognitive battery designed for efficient computerized assessment (Gur et al., 2001, 2010). The test begins by showing examinees 20 faces that they will be asked to identify later. Faces are shown in succession for an encoding period of 5 seconds each. After this initial learning period, examinees are immediately shown a series of 40 faces—20 targets and 20 distractors—and are asked to decide whether they have seen each face before by choosing 1 of 4 ordered categorical response options: 1 = "definitely yes"; 2 = "probably yes"; 3 = "probably no"; or 4 = "definitely no." For consistency with IRT's assumption of a monotonically increasing response function, we recoded responses prior to analyses so that higher valued responses always implied more accurate responses (i.e., 1 is the least correct answer, and 4 is the most correct answer for all items). Stimuli consist of black-and-white photographs of faces presented on a black background. All faces were rated as having neutral expressions and were balanced for gender and age (Gur et al., 1993, 2001). Examinees' responses and response times are recorded during test administration; however, there are no time limits during recognition testing or explicit instructions to work quickly. The PFMT takes approximately 4 minutes to administer.

Typical PFMT scoring procedures produce the number of correctly recognized targets and correctly rejected distractors, and median response times for correct responses. There are no cutoffs for qualitative interpretations of performance. In previously published studies, the reliability of accuracy scores was .75, and the reliability of time scores was .87 (Gur et al., 2010). PFMT scores have also demonstrated expected associations with demographic variables such as gender (women perform better), age (younger examinees perform better), and education (more educated examinees perform better), and its strongest intercorrelations are with measures of word memory and spatial memory (Gur et al., 2010).

Model analysis

Data were analyzed using a single-sample cross-validation. The sample was first randomly split into a training set (N = 2,118) and a validation set (N = 2,118). Exploratory psychometric analyses were used to determine the number of latent factors present in the data by examining eigenvalues (including a parallel analysis that compared observed eigenvalues to 10 randomly simulated eigenvalue sets) and factor loadings for multiple varimax-rotated (orthogonal) solutions (polychoric correlations were analyzed in both techniques). We next fit exploratory between-item multidimensional and within-item multidimensional graded response models to the data (both variations of Equation 1; see Figure 2) and compared the relative fit of each using Akaike information criterion (AIC) and Bayesian information criterion (BIC) values. AIC and BIC values penalize overparameterized models and become smaller with better fit (for the use of AIC and BIC in IRT, see de Ayala, 2009).

The best fitting exploratory model from the training data was then cross-validated in the validation data to determine whether the measurement structure would generalize across samples. Item fit was evaluated by assessing the amount of local dependence remaining between items (Chen & Thissen, 1997). Large values of local dependence [residual association (φ_c^2) > .2] imply poor model fit (see de Ayala, 2009). This is because local dependencies—unaccounted for common variance between items—can distort parameter estimates and lead to incorrect conclusions regarding the reliability and validity of test scores. After verifying fit of the chosen model in the validation data, the entire sample of 4,236 participants was pooled in order to estimate final item and examinee parameter values. Lastly, to better understand the latent dimensions themselves, we compared fitted values to individual parameter estimates derived from a signal detection theory analysis.

All analyses were conducted using the R language (R Development Core Team, 2011). Eigenvalues and factor loadings were estimated using the "psych" package for R (Revelle, 2011). Exploratory factor analyses were based on maximum likelihood estimation with orthogonal (varimax) rotations. MIRT model parameters were estimated using the "mirt" package for R (Chalmers, 2011), which uses a stochastic approximation algorithm (see Cai, 2010) to fit exploratory parameter values to data. Data missing at random are accommodated in the package within a Metropolis–Hastings Robbins–Monro algorithm. Parameter estimates are based on the logistic model with the added multiplicative constant 1.702, which produces results that are nearly identically to the normal ogive model (see Haberman, 1974). Procedures and software for estimating individual differences in the unequal-variance signal detection model are less well established. Therefore, we used R to program a Metropolis–Hastings within-Gibbs sampler for a hierarchical Bayesian representation of the model

and used approximated expected a posteriori (EAP) values as the final parameter estimates. Additional details about this procedure are provided in the Appendix.

RESULTS

Exploratory analysis

Training data

The training data were analyzed using exploratory psychometric techniques. The results reveal that the first 10 eigenvalues are all greater than one (8.03, 4.56, 1.76, 1.42, 1.31, 1.23, 1.21, 1.11, 1.08, and 1.04), which is sometimes considered to be the criterion for retention. In addition, when the eigenvalues were compared against values based on random simulations (parallel analysis), a total of 13 were sufficient to retain. Despite this, the first two eigenvalues appear to be dominant, and the amount of variance explained quickly diminishes to a nearly constant value by the third factor (20%, 11%, 4%, 4%, 3%, 3%, 3%, 3%, 3%, and 3%).

The loadings for 1-, 2-, and 3-factor exploratory solutions are presented in Table 2. The results appear to favor a 2-factor solution. Specifically,

TABLE 2
Loadings for exploratory 1-, 2-, and 3-factor solutions

		1-Factor	*2-Factor*		*3-Factor*		
Type	*No.*	λ	λ_1	λ_2	λ_1	λ_2	λ_3
Target	3	**0.40**	0.19	**0.47**	0.15	**0.50**	0.18
Target	4	**0.38**	0.16	**0.49**	0.11	**0.56**	0.16
Target	5	**0.32**	0.08	**0.52**	0.03	**0.57**	0.19
Target	6	**0.36**	0.09	**0.56**	0.04	**0.63**	0.20
Target	8	**0.14**	−0.06	**0.37**	−0.07	**0.29**	0.24
Target	9	**0.40**	0.21	**0.43**	0.21	0.30	**0.31**
Target	10	**0.21**	0.04	**0.36**	0.01	**0.37**	0.15
Target	14	**0.41**	0.14	**0.58**	0.15	0.36	**0.46**
Target	19	**0.33**	0.08	**0.53**	0.07	**0.41**	0.35
Target	20	**0.34**	0.09	**0.52**	0.08	**0.41**	0.33
Target	21	**0.49**	0.20	**0.66**	0.20	0.43	**0.50**
Target	22	**0.39**	0.11	**0.60**	0.12	0.37	**0.47**
Target	25	**0.11**	−0.08	**0.34**	−0.07	0.21	**0.27**
Target	28	**0.21**	0.00	**0.41**	0.03	0.18	**0.39**
Target	29	**0.43**	0.18	**0.56**	0.23	0.19	**0.60**
Target	31	**0.23**	0.01	**0.45**	0.03	0.19	**0.44**
Target	34	**0.14**	−0.05	**0.37**	−0.02	0.08	**0.44**
Target	35	**0.18**	−0.04	**0.42**	0.00	0.11	**0.48**
Target	36	**0.22**	−0.03	**0.49**	−0.01	0.23	**0.46**
Target	40	**–0.01**	−0.14	**0.23**	−0.10	−0.04	**0.36**
Distractor	1	**0.51**	**0.46**	0.23	**0.51**	0.01	0.31
Distractor	2	**0.20**	**0.24**	−0.01	**0.27**	−0.11	0.10
Distractor	7	**0.29**	**0.40**	−0.10	**0.44**	−0.21	0.06
Distractor	11	**0.38**	**0.44**	0.01	**0.46**	−0.07	0.09
Distractor	12	**0.45**	**0.49**	0.05	**0.51**	−0.01	0.09
Distractor	13	**0.48**	**0.54**	0.03	**0.57**	−0.03	0.09
Distractor	15	**0.45**	**0.53**	−0.01	**0.54**	−0.03	0.03
Distractor	16	**0.35**	**0.47**	−0.10	**0.49**	−0.10	−0.03
Distractor	17	**0.70**	**0.60**	0.35	**0.60**	0.28	0.22
Distractor	18	**0.68**	**0.60**	0.30	**0.59**	0.27	0.15
Distractor	23	**0.49**	**0.56**	0.00	**0.56**	0.05	−0.03
Distractor	24	**0.50**	**0.57**	0.04	**0.57**	0.05	0.01
Distractor	26	**0.53**	**0.60**	0.03	**0.57**	0.16	−0.10
Distractor	27	**0.63**	**0.65**	0.16	**0.62**	0.26	−0.02
Distractor	30	**0.30**	**0.45**	−0.16	**0.43**	0.00	−0.22
Distractor	32	**0.52**	**0.59**	0.03	**0.57**	0.16	−0.10
Distractor	33	**0.67**	**0.65**	0.20	**0.62**	0.33	−0.03
Distractor	37	**0.70**	**0.65**	0.27	**0.63**	0.34	0.05
Distractor	38	**0.51**	**0.54**	0.08	**0.51**	0.21	−0.09
Distractor	39	**0.70**	**0.66**	0.26	**0.64**	0.31	0.06

Note. λ = factor loading. Largest values shown in bold for 2- and 3-factor solutions.

2 factors maximized the number of items for which there is just one large, primary loading per item. For ease of interpretation, items in Table 2 are ordered by targets (first 20 rows; item numbers 3, 4, . . . 40) and then distractors (second 20 rows; item numbers 1, 2, . . . 39). As can be seen, this provides a clear distinction in the 2-factor solution: Factor 1 is primarily related to correct rejections, and Factor 2 is primarily related to hits. For clarity, the first of these is referred to as the distractor-loaded factor, and the second is referred to as the target-loaded factor. The 1-factor solution is difficult to interpret, as it appears to create one amalgam of factors with no consistent pattern of loadings. The 3-factor solution retains the distractor-loaded factor, but splits the target-loaded factor into two subfacets, with items tending to load onto the first of these early in the test and onto the second late in the test. The results of 4- or more factor solutions (not reported here) suggest a high number of difficult to interpret single-item factors or factors with no dominant loadings at all. Because the amount of variance accounted for by the first two factors could be readily distinguished from higher factor solutions, coupled with the interpretability of the two-factor solution and its parsimony, we decided to proceed with the analysis using the 2-factor model. We return to this issue in the Discussion.

To explore the need for nonprimary ability loadings, 2-factor M-GRMs were fit to the data to determine whether items are better modeled by within- or between-item multidimensionality (i.e., if the nonprimary ability loadings could be eliminated; center panel of Figure 2 vs. right panel of Figure 2). Both within- and between-item models converged without difficulty. The AIC and BIC values for the within-item model (162,433.30 and 163,564.90) were consistently lower than the AIC and BIC values for the between-item model (163,104.50 and 164,009.80). This suggests that within-item multidimensional structure provides a better fit than between-item multidimensional structure, even after accounting for the former's increased complexity due to a higher number of parameters that must be estimated.

Cross validation

The training data suggest that the latent structure of the PFMT is adequately fit by a 2-factor within-item M-GRM. To confirm this, the model was next applied to the validation data and was evaluated for item fit. Local dependence values are shown in the top panel of Figure 3. As can be seen, the majority of values are less than .10, and none are greater than .20. The average value was .08, which implies overall good fit. Specifically, it appears as though all major sources of common variance between items can be attributed to just two latent variables without great ill effect on the final parameter estimates.

A correlation plot (heat map) showing specific item-pair local dependencies is shown in the bottom panel of Figure 3. Areas of concentration (darker colors) are indicative of systematic model fit issues. As was done in Table 2, the items are grouped by targets (Items 3, 4, . . . , 40) and then distractors (Items 1, 2, . . . , 39). Again, it can be seen that most item-pair associations were very low. However, there does appear to be a tendency for items administrated later in the PFMT task (i.e., Items 30 through 40) to show higher local dependence values than other item pairings.

Comparison to cognitive model

Item parameter estimates derived from the entire sample are reported in Table 3, and graphical representations in the form of item vectors are shown in Figure 4 (for generation of these plots see Reckase, 2009). Panel A of Figure 4 shows item vectors plotted in a coordinate system comprised of orthogonal axes (θ_1 and θ_2), which we have called the distractor- and target-loaded factors, respectively. The angles formed between the vectors and axes convey the directions of association between the items and factors (i.e., smaller angles imply greater relative discrimination). The overall vector lengths, which are equal to the items' multidimensional discrimination parameters, convey the magnitudes of association. The origins of the vectors are related to the items' thresholds; smaller valued coordinates imply less difficulty. Because each item is associated with three thresholds, each has three vectors as well.

The item threshold parameters (d_1, d_2, and d_3) are generally positive, indicating that most items were of easy to moderate difficulty (recall that difficulty and threshold are inversely related). This is also conveyed by the high prevalence of item vectors originating in the upper left and lower left quadrants of Panel A in Figure 4. Yet, while nearly all of the d_3 parameters for targets in Table 3 are positive (d_3 item vector coordinates less than 0), d_3 parameters for distractors are mostly negative (d_3 item vector coordinates greater than 0).

The item discrimination parameter estimates (a_1 and a_2) indicate that the first latent ability (θ_1) is more strongly associated with distractors (i.e., distractor-loaded factor) and that the second latent

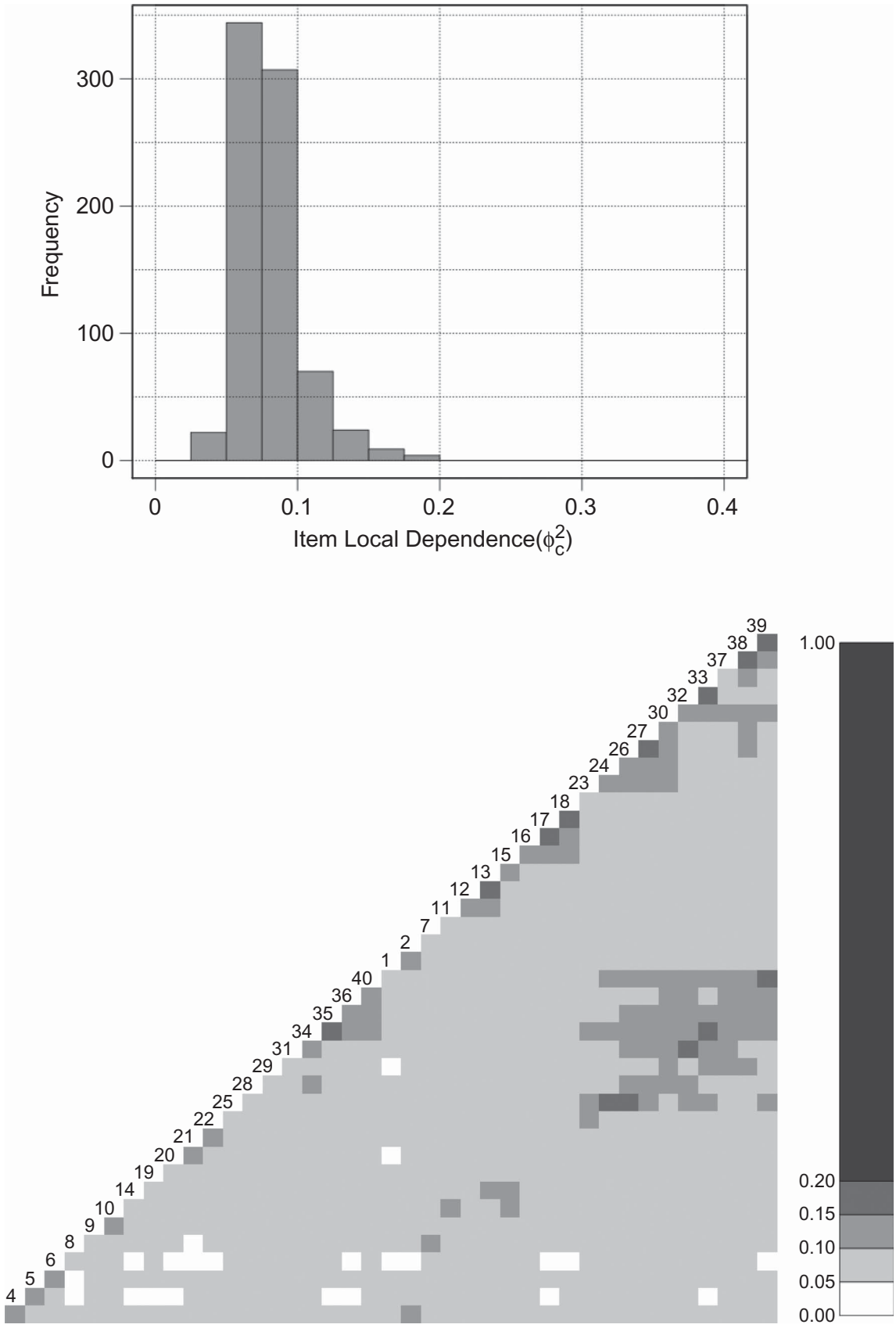

Figure 3. Top: Histogram of local dependence (φ_c^2) between items after fitting the within-item multidimensional graded response model to the data. Bottom: Heat map of local dependence (φ_c^2) between items. Darker colors imply greater local dependence.

TABLE 3
Parameter estimates for the within-item multidimensional graded response model

		Item threshold			*Item discrimination*			
Type	*No.*	d_1	d_2	d_3	a_1	a_2	a_1^*	a_2^*
Target	3	1.60	1.05	0.61	0.44	0.37	0.56	0.14
Target	4	2.09	1.48	1.02	0.48	0.47	0.64	0.22
Target	5	1.76	1.24	0.91	0.35	0.41	0.49	0.22
Target	6	2.32	1.59	1.00	0.42	0.53	0.61	0.29
Target	8	1.34	0.55	0.08	0.09	0.38	0.25	0.30
Target	9	1.56	0.99	0.65	0.44	0.39	0.57	0.15
Target	10	1.24	0.52	0.18	0.23	0.37	0.37	0.23
Target	14	2.21	1.59	1.01	0.51	0.56	0.70	0.28
Target	19	1.76	1.30	0.57	0.36	0.48	0.53	0.28
Target	20	1.67	1.07	0.62	0.40	0.51	0.58	0.28
Target	21	2.18	1.62	1.22	0.66	0.64	0.87	0.28
Target	22	1.72	1.06	0.65	0.41	0.55	0.61	0.31
Target	25	0.92	0.31	−0.07	0.13	0.37	0.28	0.28
Target	28	1.34	0.72	0.10	0.19	0.47	0.38	0.34
Target	29	1.98	1.40	1.06	0.58	0.54	0.76	0.23
Target	31	1.50	0.76	0.23	0.23	0.49	0.42	0.34
Target	34	0.98	0.40	−0.06	0.16	0.43	0.34	0.31
Target	35	1.16	0.48	0.03	0.19	0.47	0.38	0.34
Target	36	1.55	0.82	0.22	0.22	0.56	0.44	0.40
Target	40	0.71	0.06	−0.46	0.00	0.32	0.14	0.29
Distractor	1	1.74	1.00	0.14	0.69	−0.02	0.61	−0.32
Distractor	2	0.52	−0.14	−0.70	0.30	−0.15	0.20	−0.26
Distractor	7	0.54	0.13	−0.68	0.39	−0.29	0.23	−0.43
Distractor	11	0.96	0.47	−0.40	0.48	−0.20	0.34	−0.39
Distractor	12	1.24	0.73	−0.24	0.61	−0.18	0.46	−0.43
Distractor	13	1.27	0.60	−0.38	0.70	−0.21	0.54	−0.50
Distractor	15	1.22	0.78	−0.18	0.65	−0.29	0.46	−0.55
Distractor	16	0.74	0.14	−0.67	0.48	−0.35	0.28	−0.52
Distractor	17	2.98	2.31	0.86	1.20	0.08	1.11	−0.46
Distractor	18	2.58	2.02	0.91	1.12	0.04	1.03	−0.46
Distractor	23	1.41	0.87	−0.26	0.73	−0.27	0.54	−0.56
Distractor	24	1.38	0.79	−0.26	0.69	−0.29	0.50	−0.56
Distractor	26	1.65	1.11	−0.01	0.81	−0.28	0.61	−0.61
Distractor	27	2.19	1.48	0.08	1.06	−0.23	0.85	−0.67
Distractor	30	0.55	0.05	−0.85	0.41	−0.42	0.18	−0.56
Distractor	32	1.46	0.92	−0.28	0.75	−0.27	0.56	−0.57
Distractor	33	2.27	1.67	0.32	1.09	−0.12	0.93	−0.59
Distractor	37	2.02	1.52	0.38	1.09	−0.02	0.97	−0.50
Distractor	38	1.07	0.61	−0.08	0.62	−0.15	0.50	−0.41
Distractor	39	2.54	1.85	0.44	1.18	−0.07	1.03	−0.58

Note. a_1^* and a_2^* are rotated parameters.

ability (θ_2) is more strongly associated with targets (i.e., target-loaded factor). This can also be seen in Panel A of Figure 4 where the light-gray item vectors (distractors) tend to align with θ_1 and the dark-gray item vectors (targets) tend to align with θ_2. Unfortunately, this manifestation of the model is complicated by rotational indeterminacy. That is, due to the exploratory nature of the M-GRM, there is no correct orientation to the *x*- and *y*-axes (θ_1 and θ_2). Any consistent orthogonal rotation of the vector and axis values will produce the same model fit, with just the item weights changing. A proper rotation of the axes cannot be determined by the observed data alone, but instead must be "resolved" using a rotation procedure based on a theory-based criterion.

Our approach is to resolve this rotational indeterminacy using cognitive–psychometric modeling. Specifically, results from the M-GRM were correlated with individual parameter estimates derived from a specific application of the signal detection theory of recognition memory—the unequal-variance model. Correlations between the unrotated M-GRM dimensions (θ_1 and θ_2) and the estimated signal detection parameters are depicted in Panel B of Figure 4. The vectors for d',

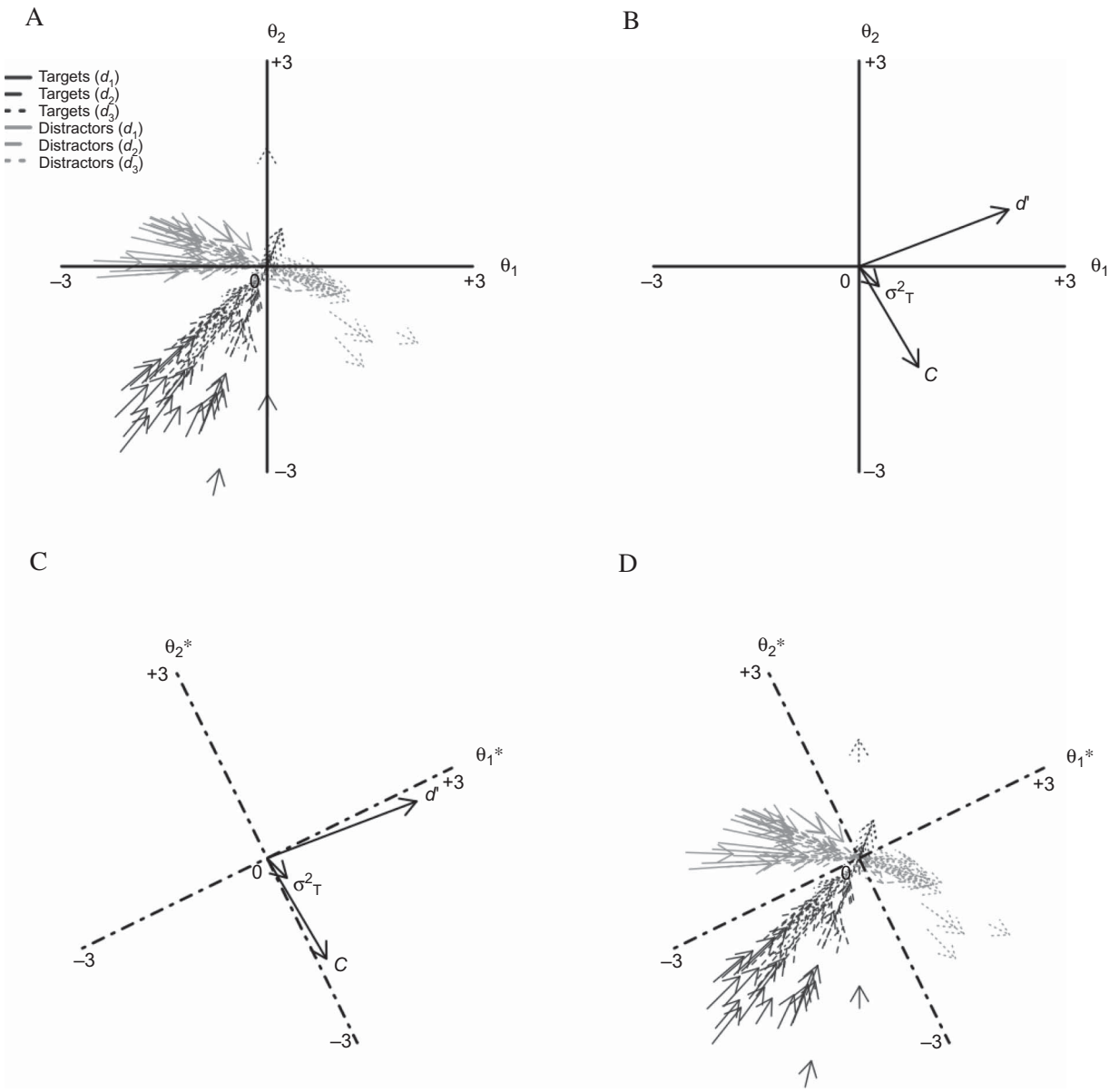

Figure 4. Unrotated (A and B) and rotated (C and D) solutions to the within-item multidimensional graded response model. A: Unrotated orientation of the axes (θ_1 and θ_2) plotted with item vectors. B: Unrotated orientation plotted with unequal variance signal detection model parameter vectors. C: Rotated orientation of the axes (θ_1* and θ_2*) plotted with unequal variance signal detection model parameter vectors. D: rotated orientation plotted with item vectors. Signal detection parameters are distance between the means of the target and distractor distributions (d'), variance of the target distribution (σ^2_T), and response bias (C).

σ^2_T, and C are plotted at angles and lengths[2] that convey the direction and magnitude of association, respectively, between each parameter and the unrotated θ_1 and θ_2 dimensions (smaller angles and longer vectors indicate stronger association). The unrotated M-GRM results provide no more than an ambiguous interpretation of the distractor- and target-loaded abilities with respect to the signal detection parameters. The dimension θ_1 is positively correlated with both d' and C. The dimension θ_2 is positively correlated with d' but negatively correlated with C. Neither ability is strongly correlated with σ^2_T, which could reflect poor estimation of the parameter. The unrotated solution was unexpected from the perspective of signal detection models of recognition memory. In particular, the latent constructs, which ought to align with memory strength and response

[2]Because the angles and vector lengths were based on the sample data, rather than the population model, the values do not maintain orthogonal geometry within the axes. Therefore, we used the average angle between each dimension to plot the vectors.

bias, should manifest themselves equally in target responses and distractor responses.

To improve clarity on this issue, we next orthogonally rotated the solution to align θ_1 with d' and θ_2 with C (an angle of approximately 26°). Specifically, by first averaging the angle of separation between θ_1 and d' with the angle of separation between θ_2 and C, and then consistently rotating all parameter estimates by this amount, results from the M-GRM model were brought into alignment with results from the signal detection model. This is similar to classic targeted rotation procedures (see Mulaik, 2010), except that the angle of rotation was determined directly from associations among estimates of ability and the signal detection parameters rather than a targeted matrix for the item parameters. The resulting dimensions after rotation along with the signal detection parameter vectors are shown in Panel C of Figure 4. It can be seen that the rotation improves interpretation of the M-GRM latent dimensions considerably. The first ability (θ_1^*) can now be regarded as an indicator of memory strength (i.e., closely aligned with d'), and the second ability (θ_2^*) can now be regarded as an indicator of response bias (i.e., closely aligned with C, although in the opposite direction). Because θ_1 and θ_2 were rotated by the same angle (i.e., orthogonal rotation), the dimensions remain orthogonal to one another.

The rotated dimensions θ_1^* and θ_2^* along with the PFMT item vectors are shown in Panel D of Figure 4. Note that θ_1^* bisects the average angle between distractor and target item vectors. This suggests that θ_1^* reflects a general ability, where higher values of θ_1^* are associated with higher probabilities of correct response for both targets and distractors. Also note that θ_2^* is positively related to targets and negatively related to distractors. This suggests that θ_2^* reflects a tradeoff, where higher values of θ_2^* are associated with higher probabilities of correct response to targets but lower probabilities of correct response to distractors, and that lower values of θ_2^* are associated with lower probabilities of correct response to targets but higher probabilities of correct response to distractors.

The rotated discrimination parameters are reported in Table 3 (a_1^* and a_2^*). Importantly, the items' overall multidimensional discriminating powers (and communalities) have not changed. Discrimination parameters for the first rotated dimension (a_1^*) are all positive and generally moderate in size. Better memory strength leads to a higher number of hits and correct rejections. Discrimination parameters for the second rotated dimension (a_2^*) are of the opposite sign for targets and distractors. That is, high θ_2^* implies bias towards targets, and low θ_2^* implies bias towards distractors. The average absolute value of a_1^* is larger than the average absolute value of a_2^*, which is more true of targets ($|M_{a1^*}| = .50$ vs. $|M_{a2^*}| = .28$) than of distractors ($|M_{a1^*}| = .60$ vs. $|M_{a2^*}| = .50$).

Measurement precision for Army STARRS

The SE_θ function for the PFMT is plotted in the top panel of Figure 5. Curvature in the SE_θ surface shows that precision in measurement varies as a function of ability. The function reaches a low point around one standard deviation below average memory strength and average response bias (the coordinate $\theta_1^* = -1.0$, $\theta_2^* = -1.3$). Notably, there is one "valley" of SE_θ running diagonally[3] from about ($\theta_1^* = -4.0$, $\theta_2^* = -4.0$) to ($\theta_1^* = 4.0$, $\theta_2^* = 4.0$). High points in the SE_θ surface fall in regions where memory strength is very high, and response bias is very low (e.g., $\theta_1^* = 4.0$, $\theta_2^* = -4.0$) or where memory strength is very low, and response bias is very high (e.g., $\theta_1^* = -4.0$, $\theta_2^* = 4.0$).

The bottom panel of Figure 5 plots the bivariate distribution of ability estimates in the Army STARRS sample. The highest density of ability occurs at ($\theta_1^* = 0$, $\theta_2^* = 0$), with both memory strength and response bias symmetrically distributed about the point. By visually lining-up the ability distribution in Figure 5 with the SE_θ function in Figure 5, it can be seen that there is a close match. That is, low values of SE_θ tend to occur near dense ability regions. However, the SE_θ function does not appear to be optimized for this specific sample. The low point in the SE_θ function suggests that ability estimates are most precise for slightly less able individuals. Indeed, estimated reliability in the current sample is .61; however, if we consider only those Soldiers with ability estimates falling within one standard deviation of the low point in the SE_θ surface ($\theta_1^* = -1.0$ and $\theta_2^* = -1.3$), reliability jumps to .71. If we consider only those Soldiers with ability estimates falling within one standard deviation of an arbitrary high point in the SE_θ surface (say $\theta_1^* = 1.0$ and $\theta_2^* = -1.0$),

[3] SE_θ functions are difficult to summarize for within-item multidimensional models. Whereas changes in ability for unidimensional models can occur in just one dimension, changes in ability for multidimensional models can occur in multiple dimensions. Therefore, in order to determine how SE_θ changes as a function of ability, a gradient describing the direction of movement within the ability plane must first be established. For simplicity, we aligned this gradient with the direction of steepest descent. In practice, observed SE_θ values are likely to vary depending on changes in this angle.

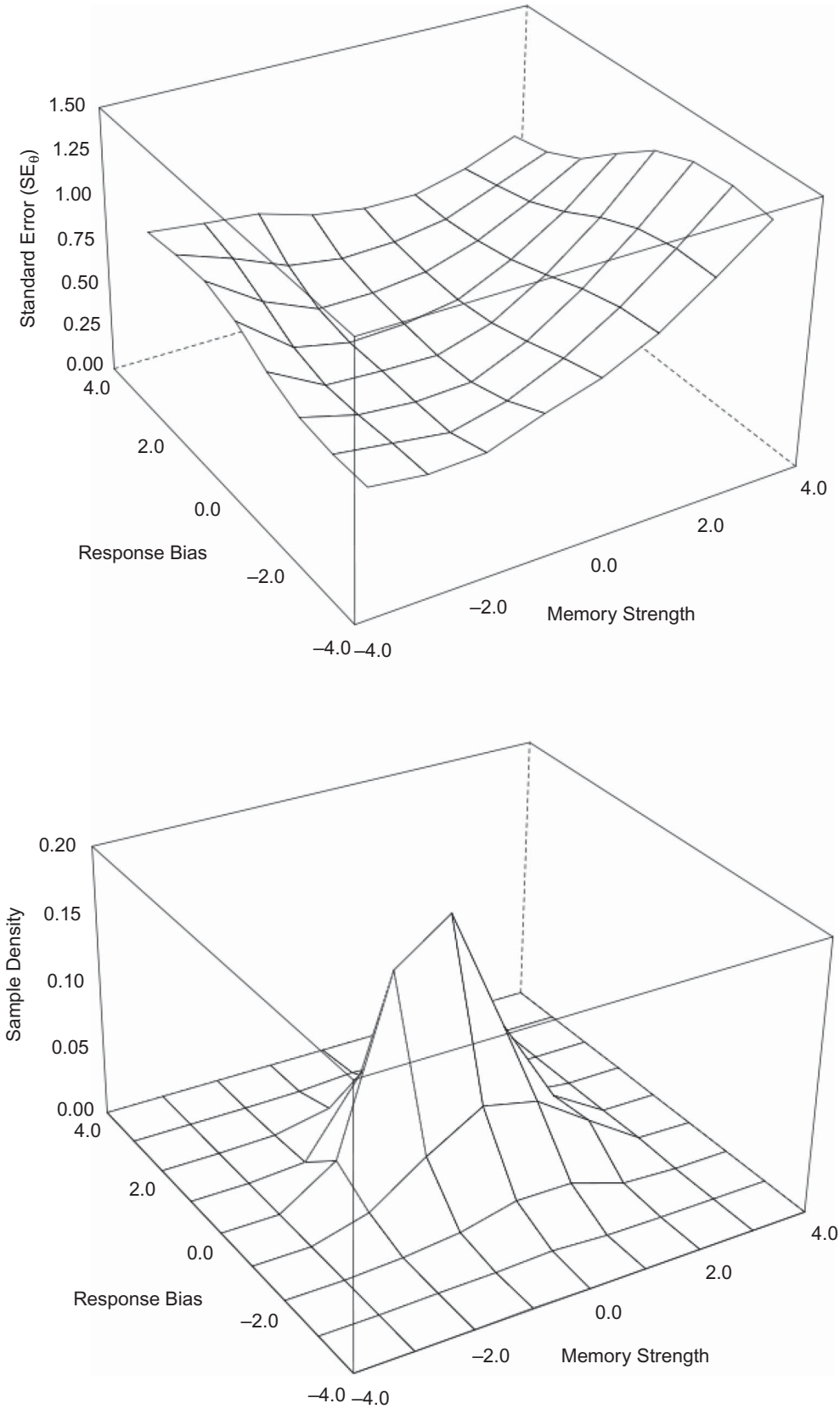

Figure 5. Top panel: standard error function for the rotated solution to the within-item multidimensional graded response model. Bottom panel: multidimensional distribution of memory strength and response bias in the Army STARRS sample.

reliability drops to .42 and would be much lower had we chosen more extreme values.

DISCUSSION

Parallel psychometric and cognitive modeling analyses were used to better understand the measurement properties of the PFMT as a component of the neuropsychological battery being administrated to Soldiers in Army STARRS. Item response theory (M-GRM model) was used to confirm the number of latent cognitive abilities measured by the instrument. Then, cognitive theory (unequal-variance signal detection model) was used to apply psychological meaning to those abilities. Specifically, these analyses supported the construct validity of independent neurocognitive markers of memory strength (d') and response bias (C). We then investigated the measurement precision of PFMT test scores and found adequate precision overall, but higher precision in low-average ability regions. Thus, PFMT scores were most precise for examinees with low-average memory strength. The results suggest that although the PFMT can provide useful measurements across a wide range of memory abilities, it is particularly useful for identifying differences among impaired individuals.

Cognitive–psychometric methodology helped explain unknown aspects of the test's latent factor structure using experimentally derived cognitive architecture. The approach provided a synthesis of statistical modeling and cognitive theory. Although the modeling procedures were complex in comparison to conventional approaches, the final measurements had a priori psychological meaning and did not require a posteriori adjustments (e.g., multiple regression or pattern analysis) to be made interpretable. In addition, the modeling procedures allowed us to examine standard error as a function of cognitive ability. This facilitated a more nuanced inspection of the PFMT's psychometric properties and allowed us to determine specific levels of memory associated with relatively better or worse measurement precision.

Penn Face Memory Test model structure

The within-item M-GRM (bottom panel of Figure 1 and right panel of Figure 2) is considered to be a relatively complex measurement structure. Within-item multidimensional models require greater effort to estimate and to interpret than do between-item models (McDonald, 2000). A common solution to such complexity is to force between-item (simple) structure by allowing factors to correlate with each other. In Panel A of Figure 4, for example, it is possible to obliquely rotate the x- and y-axes so that the target and distractor items primarily align with just one dimension. That is, it is possible to avoid models where items are indicators of multiple cognitive constructs by assuming that the constructs are highly correlated. Although such a solution is methodologically simpler, it fails to agree with cognitive theories suggesting that target and distractor responses should be similarly determined by two underlying constructs. A more psychologically meaningful, albeit complex, solution was to interpret the psychometric model (within-item M-GRM) from a cognitive perspective (i.e., targeted rotation using unequal-variance signal detection model parameters). This methodology produced clear psychological interpretations for the latent abilities—one representing memory strength and one representing response bias.

Aligning the exploratory MIRT results with a well-validated cognitive model helps to establish the PFMT's construct representation—that is, how test scores (item responses) are determined by the properties of items, tasks, and underlying cognitive processes (see Embretson, 1983, 2010). In our analyses, for example, the function of the first rotated ability ($\theta_1{}^*$) was shown to be consistent with the construct representation of memory strength, and the function of the second rotated ability ($\theta_2{}^*$) was shown to be consistent with the construct representation of response bias. This stringent form of validity is supported by cognitive–psychometric modeling methodology (see Batchelder, 2010). Long recognized as the ideal synthesis of theories from experimental and correlational psychology (Cronbach, 1957), this methodology is rapidly expanding its applications (e.g., Brandt, 2007; Brown, Turner, Mano, Bolden, & Thomas, 2012; Embretson, & Gorin, 2001; Townsend & Neufeld, 2010) and has the potential to improve neuropsychological assessment, particularly in an age of computerized measurement.

The model applied in the current study is regarded as a preliminary attempt to build a cognitive–psychometric architecture for neurocognitive measures of episodic memory. Although the model was made consistent with basic results from signal detection theory (a descriptive cognitive model), it remains unclear how specific cognitive processes, as well as other commonly studied cognitive phenomena (e.g., list length), can be accommodated by the measurement structure. Whether the cognitive–psychometric model described in this paper needs to be replaced by a synthesis of psychometric theory with a more explanatory

cognitive theory of episodic memory (see Clark & Gronlund, 1996, for examples), or can be amended to handle the inherent complexity in episodic memory testing, is unknown. Obvious limitations in the applied model are that changing the number of study items (faces) and the time between study and test phases will affect item difficulty. Also, the process of face recognition itself could affect memory strength for other stored but not yet tested faces. Consistent with these concerns, our results suggest that in addition to the two dominant factors underlying test performance (memory strength and response bias), nearly a dozen smaller factors were present in the data. Our analysis of local dependence (Figure 3) suggests that most of this residual common variance is found among items administrated later in the PFMT task. We assume that this variance reflects the impact of unaccounted for cognitive processes (e.g., learning to learn, network activation, etc.), but this hypothesis was not tested.

If our assumption is accurate, a more comprehensive model comprising structures for both cognitive abilities and cognitive processes is required to accurately account for episodic memory task performance. For example, there is some agreement in the field of cognitive modeling that memory strength can be further deconstructed into the additive (see Wixted, 2007) and/or complementary (see Onyper, Zhang, & Howard, 2010; Yonelinas, 2002) effects of familiarity and recollection. Whereas familiarity is characterized by the weak sense of having experienced some past event, recollection conveys clear and unequivocal memory probably related to memory search. Imaging studies have even suggested that these processes rely on distinct neuroanatomical structures (Davachi & Wagner, 2002; Diana, Yonelinas, & Ranganath, 2007; Kim & Cabeza, 2007; Wheeler & Buckner, 2003), though perhaps not in a clear and unambiguous manner (Wixted & Squire, 2011).

Integrative frameworks for such findings are offered by computational global memory models (for a review, see Clark & Gronlund, 1996). Most of these assume that memories, which are represented parametrically as associative networks of item and context information, are accessed through a process combining elements of global matching and interactive cueing. Recollection equates to a specific match between a test item and a stored memory; familiarity equates to the marginal associative strength between a test item and the complete stored network of memories (e.g., Gronlund & Shiffrin, 1986; Raaijmakers & Shiffrin, 1980). These computational models are unique in that they are designed to explain underlying memory structures and processes at a level of specificity not feasible with more universally applicable models (e.g., generalized linear model). Further research is needed to determine whether computational models can be given a cognitive–psychometric framework.

Measurement properties of the PFMT and implications for Army STARRS

Standard error is determined by item and examinee characteristics. Specifically, measurement precision is optimized when item difficulty is closely matched with ability and when item discrimination is high. This can be visualized in Panel D of Figure 4. The area of the two-dimensional ability space (θ_1* by θ_2*) covered by the dark- and light-gray arrows (item vectors representing combined item difficulty and discrimination characteristics) is the area of ability where precise measurement occurs. This implies that the PFMT should be most precise when measuring low-average levels of memory strength (θ_1*) and response bias (θ_2*), which can also be observed in the standard error function displayed in the top panel of Figure 5. It is noteworthy that PFMT distractor items and PFMT target items contribute towards precision of measurement at complementary angles within the multidimensional ability space. That is, hits and correct rejections contribute unique value to the assessment of episodic memory and should not be regarded as interchangeable data.

Most of the PFMT items were estimated to have low difficulties (inverse of d_1, d_2, and d_3), which resulted in the test's standard error function reaching a minimum (maximum measurement precision) in the low-average range of ability. In addition, the difficulties of specific PFMT response options (i.e., "definitely yes," "probably yes," "probably no," or "definitely no") were slightly disproportionate for targets and distractors. Targets were easier than distractors. This implies that distractors contributed relatively greater measurement precision in the high-average range of ability. All things being equal, adding distractor items to the PFMT should improve measurement precision for nonimpaired examinees, while adding target items should improve measurement precision for impaired examinees, as the adjustment should make the test more or less difficult, respectively.

Most of the PFMT items were estimated to have moderately sized discrimination parameters, but memory strength values (a_1*) were generally larger than response bias values (a_2*). Thus, memory strength was measured more precisely than

response bias. As a result, variation in memory strength had a greater effect on standard error than did variation in response bias. Although somewhat difficult to discern in Figure 5, marginal variation in the memory strength axis is greater than marginal variation in the response bias axis. The surface does not form a consistent "U" shape. An implication is that, because the variances of both latent dimensions were standardized, memory strength was a greater relative determinant of response performance than was response bias, which could reflect relatively less true variance in Army Soldiers' response biases than in Army Soldiers' memory strengths. Practically, PFMT users can consider the instrument a better indicator of memory strength than response bias, which is consistent with the purpose of the test.

Overall, analyses of the PFMT suggest that standard error values were low for the majority of Army STARRS participants' ability estimates. The reliability of all estimates in the sample was .61. Thus, it can be argued that the PFMT approaches the general standard for adequate measurement precision for the Army STARRS sample as a whole. However, the reliability is much higher (.71) if we consider only those estimates in the low-average range of ability. This confirms that the PFMT's measurement precision is optimized for examinees with low-average ability. Item difficulty would have to be raised (item thresholds lowered) by approximately one standard deviation in order for measurement precision to peak in the center of the Army STARRS ability distribution. However, doing so may not be desirable in a study of psychopathology, where the impaired end of the ability spectrum is the primary focus of predictive modeling. Because Army STARRS seeks to identify potential risk factors, such as below-average cognitive functioning, the results generally support maintaining the test in its current instantiation. The capability to determine this is a benefit of modeling both SE_θ and reliability.

Two aspects of the PFMT likely contribute to its favorable measurement precision: (a) Items are generally of moderate difficulty; and (b) examinees are asked to respond using ordered categorical response options. Items with moderate difficulties are optimal for assessment of normative samples. PFMT items, which are somewhat geared towards clinical populations, can be characterized as being easy, but not too easy. Developers of fixed-length tests, particularly tests administered in time-limited settings, must choose a range of ability for which items will provide maximum precision. For the PFMT, even though ability measurements are most precise for examinees in the low-average range, examinees in the above-average range are challenged enough by the content to avoid dramatic ceiling effects in the Army STARRS sample. Standard error values are also lowered by the PFMT's rating scale format. Modeling intermediate response options (i.e., "probably") tends to minimize standard error in the bodies of ability distributions; extreme response options (i.e., "definitely") tend to minimize standard error near the tails. The PFMT's polytomous format (ordered categorical response options) has the effect of creating a pseudocontinuous response scale, which helps to overcome some of the limitations due to categorical response data.

It is important to note that an accurate characterization of standard error goes beyond evaluations of a test's measurement properties. Standard error can also be used to improve predictive modeling, as is planned for risk and resilience variables in Army STARRS. A key tenant of structural equation modeling (see Bollen, 1989) is that a properly specified measurement model can alleviate some of the problems that arise from regression with latent variables. Most prominently, regression coefficients become attenuated (shrink towards zero) as measurement error grows larger. Because standard error is a property of ability score estimates, it should be examined in predictive modeling to ensure that poorly measured examinees will not disproportionately impact analyses of a test's predictive value.

Limitations

A limitation of all latent variable studies is the inability to unequivocally define measured constructs. Although the rotated dimensions appear to represent memory strength and response bias, there is no guarantee that the PFMT measures these two latent variables. Evidence of nomothetic span (convergent and discriminant validity), much of which has been collected for the PFMT (e.g., Gur et al., 2001; Gur et al., 1997; Gur et al., 2012; Gur et al., 2010), further supports the test's construct validity. In addition, our choice to align latent ability estimates from the M-GRM with parameter values from the unequal-variance signal detection model was based on an a priori chosen theoretical framework for interpreting response data. As with all models, it is not possible to prove that the rotation is correct or is a true reflection of reality. Indeed, any rotation of the parameter space (e.g., based on simple structure, targeted item parameters, etc.) would result in the same model fit. By showing how the rotation solution can be linked to a particular

theory, however, we refocus the discussion of factor rotation strategy from mathematical criteria (varimax, promax, etc.) to criteria for adequate theory construction.

We noted that the response profiles of approximately 8% of Army Soldiers in the initial sample were flagged as errant and were therefore not included in the analyses. The exclusions may have eliminated some valid response profiles resulting in overestimates of the PFMT's model fit and measurement precision. It was also noted that demographic characteristics of Soldiers with data present were not identical to demographic characteristics of Soldiers with data absent or excluded. Disparities in gender, education, and race and ethnicity, though small, were observed between the two groups. These differences threaten the validity of findings from this study. However, perhaps more concerning than difference between Soldiers tested and Soldiers not tested is the more general possibility of group biases. Specifically, it is possible that differential item functioning between demographically diverse groups of examinees, even within the same calibration sample, led to poor estimation of some parameter values. A potential solution is to divide the current sample into subsets based on demographic characteristics and then investigate differential item functioning between groups. Such analyses, while valuable, would add considerable length to the current paper. They will be pursued in future studies.

Finally, unlike a true cross-validation, where independent samples are drawn from the same population, our analyses were based on independent samples drawn from the same sample. This implies that our results were cross-validated to the sample but not population of Army Soldiers.

CONCLUSIONS

The PFMT was found to measure visual memory strength and response bias with adequate precision for the majority of Soldiers participating in Army STARRS. The results are promising and suggest that the test has the potential to provide useful measurements of Army Soldiers' cognitive functioning. Although the psychometric model applied to the PFMT in the current study likely does not account for all relevant sources of task and item variance, it does provide a basis for future cognitive–psychometric modeling of the instrument.

Original manuscript received 6 May 2012
Revised manuscript accepted 27 December 2012
First published online 30 January 2013

REFERENCES

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Army STARRS. (2011). *National Institute of Mental Health Strategic Plan*. Army STARRS Home. Retrieved October 11, 2011, from http://www.armystarrs.org/

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 47–89). New York, NY: Academic Press.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 71–93). Washington, DC: American Psychological Association.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Brandt, M. (2007). Bridging the gap between measurement models and theories of human memory. *Journal of Psychology*, *215*, 72–85.

Brown, G. G., Turner, T., Mano, Q. R., Bolden, K., & Thomas, M. L. (2012). *Experimental manipulation of working memory model parameters: An exercise in construct validity*. Manuscript submitted for publication.

Cabeza, R., & Kingstone, A. (Eds.). (2006). *Handbook of functional neuroimaging of cognition* (2nd ed.). Cambridge, MA: MIT Press.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*, 33–57.

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

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Chalmers, P. (2011). *mirt: Multidimensional item response theory*. Retrieved from http://personality-project.org/r/psych.manual.pdf

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, *3*, 37–60.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*, 163–191.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Davachi, L., & Wagner, A. (2002). Hippocampal contributions to episodic encoding: Insights from relational

and item-based learning. *Journal of Neurophysiology*, *88*, 982–990.
de Ayala, R. J. (1994). The influence of dimensionality on the graded response model. *Applied Psychological Measurement*, *18*, 155–170.
de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *The California Verbal Learning Test* (2nd ed.). San Antonio, TX: The Psychological Corporation.
Diana, R. A. Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: A three-component model. *Trends in Cognitive Sciences*, *11*, 379–386.
Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.
Embretson, S. E. (2010). Cognitive design systems: A structural modeling approach applied to developing a spatial ability test. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 274–273). Washington, DC : American Psychological Association.
Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343–368.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.
Gronlund, S. D., & Shiffrin, R. M. (1986). Retrieval strategies in recall of natural categories and categorized lists. *Journal of Experimental Psychology*, *12*, 550–561.
Gur, R. C., Erwin, R. J., & Gur, R. E. (1992). Neurobehavioral probes for physiologic neuroimaging studies. *Archives of General Psychiatry*, *49*, 409–414.
Gur, R. C., Jaggi, J. L., Ragland, J. D., Resnick, S. M., Shtasel, D., Muenz, L., et al. (1993). Effects of memory processing on regional brain activation: Cerebral blood flow in normal subjects. *The International Journal of Neuroscience*, *72*, 31–44.
Gur, R. E., Jaggi, J. L., Shtasel, D. L., & Ragland, J. D. (1994). Cerebral blood flow in schizophrenia: Effects of memory processing on regional activation. *Biological Psychiatry*, *35*, 3–15.
Gur, R. C., Ragland, J. D., Moberg, P. J., Turner, T. H., Bilker, W. B., Kohler, C., et al. (2001). Computerized neurocognitive scanning: I. Methodology and validation in healthy people. *Neuropsychopharmacology*, *25*, 766–776.
Gur, R. C., Ragland, J. D., Mozley, L. H., Mozley, P. D., Smith, R., Alavi, A., et al. (1997). Lateralized changes in regional cerebral blood flow during performance of verbal and facial recognition tasks: Correlations with performance and "effort." *Brain and Cognition*, *33*, 388–414.
Gur, R. C., Richard, J., Calkins, M. E., Chiavacci, R., Hansen, J. A., Bilker, W. B., . . . Gur, R. E. (2012). Age group and sex differences in performance on a computerized neurocognitive battery in children age 8–21. *Neuropsychology*, *26*, 251–265.
Gur, R. C., Richard, J., Hughett, P., Calkins, M. E., Macy, L., Bilker, W. B., . . . Gur, R. E. (2010). A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: Standardization and initial construct validation. *Journal of Neuroscience Methods*, *187*, 254–262.
Haberman, S. J. (1974). *The analysis of frequency data*. Chicago, IL: University of Chicago Press.
Haynes, S. N., Smith, G. T., & Hunsley, J. D. (2011). *Scientific foundations of clinical assessment*. New York, NY: Routledge.
Hoerger, M., Quirk, S. W., & Weed, N. C. (2011). Development and validation of the Delaying Gratification Inventory. *Psychological Assessment*, *23*, 725–738.
Ingham, J. (1970). Individual differences in signal detection. *Acta Psychologica*, *34*, 39–50.
Kelley, W., Miezin, F., McDermott, K., Buckner, R., Raichle, M., Cohen, N., . . . Petersen, S. E. (1998). Hemispheric specialization in human dorsal frontal cortex and medial temporal lobe for verbal and nonverbal memory encoding. *Neuron*, *20*, 927–936.
Kim, H., & Cabeza, R. (2007). Differential contributions of prefrontal, medial temporal, and sensory-perceptual regions to true and false memory formation. *Cerebral Cortex*, *17*, 2143–2150.
Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, H. J., & Fischer, J. S. (2004). *Neuropsychological assessment* (4th ed.). New York, NY: Oxford University Press.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement*, *3*, 635–694.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Magnusson, D. (1966). *Test theory*. Reading, MA: Addison-Wesley.
McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.
McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*, 99–114.
Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299.
Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal-detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*, 858–865.
Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement. Special Issue: Polytomous Item Response Theory*, *19*, 73–90.
Onyper, S. V., Zhang, Y. X., & Howard, M. W. (2010). Some-or-none recollection: Evidence from item and source memory. *Journal of Experimental Psychology*, *139*, 341–364.
R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 207–262). New York, NY: Academic Press.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 25–46.

Revelle, W. (2011). *psych: Procedures for personality and psychological research*. Retrieved from http://personality-project.org/r/psych.manual.pdf

Samejima, F. (1997). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer.

Snodgrass, J., & Corwin, J. (1988). Pragmatics of measuring recognition memory—Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1–25.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, *18*, 291–307.

Townsend, J. T., & Neufeld, R. W. J. (2010). Introduction to special issue on contributions of mathematical psychology to clinical science and assessment. *Journal of Mathematical Psychology*, *54*, 1–4.

U.S. Army. (2010). *Army health promotion, risk reduction, and suicide prevention: Report 2010*. Retrieved October 11, 2011, from http://csf.army.mil/downloads/HP-RR-SPReport2010.pdf

Wheeler, M., & Buckner, R. (2003). Functional dissociation among components of remembering: Control, perceived oldness, and content. *Journal of Neuroscience*, *23*, 3869–3880.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York, NY: Oxford University Press.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*, 152–176.

Wixted, J. T., & Squire, L. R. (2011). The medial temporal lobe and the attributes of memory. *Trends in Cognitive Sciences*, *15*, 210–217.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years research. *Journal of Memory & Language*, *46*, 441–517.

APPENDIX

MODELING PROCEDURES

Because the densities of the ordered categorical response options were estimated to be slightly disproportionate for targets and distractors, an unequal-variance version of the signal detection model was used to model the data. The unequal-variance signal detection model is not identifiable from the observed data alone. Therefore, the mean of the distractor distribution was fixed to 0, and the variance was fixed to 1 for all examinees. The mean of the target distribution for each examinee was set equal to estimates of d', and the variance (σ^2_T) was set equal to 1 plus estimates of σ^2 (see below). The middle criterion for each examinee was found by adding estimates of bias (C) to the midpoint of estimates of d'. Individual differences in the placement of the upper and lower criteria were not uniquely estimated, but were instead fixed by adding and subtracting 0.5, respectively, to each examinee's middle criterion estimate. The value of 0.5 was found to provide good fit for data aggregated at the group level. All free parameters (d', σ^2, and C) were estimated using a hierarchical Bayesian approach with a Metropolis–Hastings within-Gibbs sampler (see Gelman, Carlin, Stern, & Rubin, 2004) written in R. The chain was thinned by every 10th draw and was run for 3,000 iterations after burn-in. We used informative prior distributions to improve parameter recovery. The priors were specified as follows: $d' \sim$ normal(1.5, 0.25); $\sigma^2 \sim$ inv-gamma(5,1); and $C \sim$ normal(0, 0.25). The means of the d' and C prior distributions were based on preliminary analyses fitted at the group level. The shape and scale parameters of the inverse gamma prior result in a prior mean of 0.25 for σ^2 and a prior mean of 1.25 for σ^2_T. This value is consistent with the commonly reported ratio of distractor to target variance of 0.8 (i.e., $\sigma^2_F / \sigma^2_T = 1.00/1.25$; see Mickes, Wixted, & Wais, 2007).