Psychological Assessment
2015, Vol. 27, No. 3, 955–964

© 2015 American Psychological Association
1040-3590/15/$12.00 http://dx.doi.org/10.1037/pas0000102

# Development of an Abbreviated Form of the Penn Line Orientation Test Using Large Samples and Computerized Adaptive Test Simulation

Tyler M. Moore
Perelman School of Medicine, University of Pennsylvania

J. Cobb Scott
Perelman School of Medicine, University of Pennsylvania and Philadelphia VA Medical Center, Philadelphia, Pennsylvania

Steven P. Reise
University of California–Los Angeles

Allison M. Port, Chad T. Jackson, Kosha Ruparel, Adam P. Savitt, and Raquel E. Gur
Perelman School of Medicine, University of Pennsylvania

Ruben C. Gur
Perelman School of Medicine, University of Pennsylvania and Philadelphia VA Medical Center, Philadelphia, Pennsylvania

Visuospatial processing is a commonly assessed neurocognitive domain with deficits linked to dysfunction in right posterior regions of the brain. With the growth of large-scale clinical research studies, there is an increased need for efficient and scalable assessments of neurocognition, including visuospatial processing. The purpose of the current study was to use a novel method that combines item response theory (IRT) and computerized adaptive testing (CAT) approaches to create an abbreviated form of the computerized Penn Line Orientation Test (PLOT). The 24-item PLOT was administered to 8,498 youths (aged 8–21 years) as part of the Philadelphia Neurodevelopmental Cohort study and, by Web-based data collection, in an independent sample of 4,593 adults from Great Britain as part of a TV documentary. IRT-based CAT simulations were used to select the best PLOT items for an abbreviated form by performing separate simulations in each group and choosing only items that were selected as useful (i.e., high item discrimination and in the appropriate difficulty range) in at least 1 of the simulations. Fifteen items were chosen for the final, short form of the PLOT, indicating substantial agreement among the models in how they evaluated each item's usefulness. Moreover, this abbreviated version performed comparably to the full version in tests of sensitivity to age and sex effects. This abbreviated version of the PLOT cuts administration time by 50% without detectable loss of information, which points to its feasibility for large-scale clinical and genomic studies.

*Keywords:* psychometrics, Penn Computerized Neurocognitive Battery, item response theory, computerized adaptive testing, line orientation test

Visuospatial processing is a core cognitive skill linked to posterior cortical function, with neuroimaging and lesion studies providing evidence of right-sided specificity (e.g., Benton, Varney, & Hamsher, 1978; Gur et al., 1982, 2000; Hannay et al., 1987; Trahan, 1998; Tranel, Vianna, Manzel, Damasio, & Grabowski, 2009). Accurate assessment of visuospatial functioning is an integral part of neurological, neuropsychological, and neuropsychiatric research and practice (e.g., Rabin, Barr, & Burton, 2005).

Because of the recent growth in large-scale clinical and genomic studies, there are increasing demands for efficient, valid, and scalable assessments of neurocognitive performance that can be used as endophenotypes of illness (Insel & Cuthbert, 2009). The Penn Computerized Neurocognitive Battery (CNB; http://www.med.upenn.edu/bbl/) was designed to address this need (Gur et al., 2001, 2010) by offering a set of "neurobehavioral probes" (Gur, Erwin, & Gur, 1992) validated with functional neuroimaging

This article was published Online First March 30, 2015.

Tyler M. Moore, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania; J. Cobb Scott, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania and VISN4 Mental Illness Research, Education, and Clinical Center, Philadelphia VA Medical Center, Philadelphia, Pennsylvania; Steven P. Reise, Department of Psychology, University of California–Los Angeles; Allison M. Port, Chad T. Jackson, Kosha Ruparel, Adam P. Savitt, and Raquel E. Gur, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania; Ruben C. Gur, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania and VISN4 Mental Illness Research, Education, and Clinical Center, Philadelphia VA Medical Center, Philadelphia, Pennsylvania.

This work was supported by National Institute of Mental Health grants MH089983, MH019112, and MH096891. J. Cobb Scott's contribution was supported by Department of Veterans Affairs Career Development Award 1IK2CX000772.

Correspondence concerning this article should be addressed to Ruben C. Gur, Brain Behavior Laboratory, Perelman School of Medicine, University of Pennsylvania, 3400 Spruce Street, 10th Floor Gates Pavilion, Philadelphia, PA 19104. E-mail: gur@upenn.edu

(Roalf et al., 2014) and with established psychometric properties (Moore, Reise, Gur, Hakonarson, & Gur, 2014). Although the tests that compose the CNB were often adapted from traditional neuropsychological assessments, they also have the advantage of being validated with functional neuroimaging as reflecting the recruitment of specific brain systems (e.g., Gur et al., 2000; Roalf et al., 2013, 2014; Satterthwaite et al., 2013), making them particularly useful as biomarkers of brain dysfunction (Gur et al., 2012). To this end, the CNB has been deployed in multiple large-scale genomic, neurobehavioral, and treatment studies (Aliyu et al., 2006; Almasy et al., 2008; Grant, Huh, Perivoliotis, Stolar, & Beck, 2012; Greenwood et al., 2007; Gur et al., 2001, 2012; R. E. Gur, Calkins, et al., 2007; Gur, Nimgaonkar, et al., 2007; Thomas et al., 2013).

The Penn Line Orientation Test (PLOT) is included in the CNB to assess visuospatial processing with minimal motor or language demands. In the PLOT, two line segments are presented on the screen, and participants are asked to rotate a movable line so that it is parallel to the fixed line. To rotate the line, the participant clicks repeatedly on one of two buttons that rotate the line clockwise or counterclockwise for each click. The number of degrees of rotation for each click varies from 3°, 6°, or 9°, producing increased precision demand, and hence difficulty, with lower degrees of rotation per click. In each trial, the location of the lines relative to one another varies, but the distance between the centers remains constant. The length of the movable line also varies among three lengths in different trials, but the length of the fixed line remains constant. There are a total of 24 trials in the test. The final orientations of the lines, as well as the efficiency of the path used to reach that orientation, are recorded. The test exhibits adequate psychometric properties (Gur et al., 2001, 2010; Moore et al., 2014) and has been used in large-scale genomic studies to examine associations with psychiatric disorders and brain structure and function (Gur et al., 2012; Iannacone et al., 2014; Van Essen et al., 2012).

Notwithstanding its strengths, the full-length PLOT is time-consuming to administer, and efficiency is increasingly being required in research and clinical assessments, especially in large-scale studies. In addition, although the PLOT has adequate psychometric properties, it is possible that some individual items provide a poorer assessment of the ability underlying performance on the PLOT than other items, and identifying such items would yield opportunities for increased efficiency in assessment. Item response theory (IRT) offers the methodology to improve instruments by incorporating information regarding how well each item discriminates among different levels of the underlying ability (i.e., item discrimination); how difficult each item is; and, although not relevant to the present case, the likelihood of guessing a correct answer on an item. Using IRT to construct more efficient versions of instruments may help avoid the unreliability of scores and inadequate structural validity often encountered when short forms are constructed with alternative methods, such as using odd-even splits of items (e.g., Spencer et al., 2013). Moreover, combining IRT with computerized adaptive testing (CAT) techniques, which tailor difficulty levels to specific test-takers based on their general abilities, can help further shorten administration time by avoiding administration of items that do not offer valuable information about specific individuals (see Segall, 2005; Weiss & Kingsbury, 1984).

This study describes the creation of a short form of the PLOT using a novel combination of IRT and CAT techniques in a large-scale sample of youth, the Philadelphia Neurodevelopmental Cohort (PNC; Calkins et al., 2014; Merikangas et al., in press; Satterthwaite et al., 2014). We also provide confirmation of these IRT models in an independent sample of adults from the United Kingdom, who took the test on the Web as part of a TV documentary, and we examine the consistency of this abbreviated form with previous literature in detecting age and sex differences in visuospatial processing.

## Method

### Participants and Settings

This study includes two independent samples. The first sample comprises 8,498 youths ages 8–21 years (51% female; 57% Caucasian; mean age = 13.4 years) who were administered a battery of neurocognitive tests as part of their participation in the National Institute of Mental Health-funded PNC study from November 2009 to October 2011 (see Calkins et al., 2014, and Satterthwaite et al., 2014 for greater detail on this cohort, including the recruitment and sampling design and Merikangas et al., in press, for information on comorbidity, sociodemographic characteristics, and epidemiologic comparability to other samples). Participant inclusion criteria were (a) being able to provide informed consent; (b) proficiency in English; and (c) being physically and cognitively capable of participating in neurocognitive and psychiatric assessments. Participants with disorders that impaired motility or cognition, including intellectual disability, significant paresis, pervasive developmental disorders, or intracranial lesions, were excluded. These exclusion criteria were intentionally liberal to recruit a representative sample of youth from the greater Philadelphia area (see Calkins et al., 2014). Participants and their guardians (for participants <18 years old) provided written informed consent or assent, and the institutional review boards at the University of Pennsylvania and Children's Hospital of Pennsylvania approved the protocol.

The second sample included 4,593 TV viewers (77% female; 91% Caucasian; mean age = 34.1 years) in the United Kingdom who were invited to take the test over the Web after a TV documentary on sex differences was aired on BBC. Participants were shown a brief textual explanation of the task itself and the reason for its administration. They were given the option either to "participate" in the described research, in which case they were taken to the demographic questionnaire and task itself, or to decline to participate, in which case they were routed back to the TV program's website. Participants younger than 18 years of age were excluded because it was not feasible to obtain participant's assent and parental consent. This protocol was approved by the University of Pennsylvania Institutional Review Board.

### Assessments

As described previously (Gur et al., 2012; Moore et al., 2014), all PNC participants were administered the CNB, which consisted of 14 tests measuring a broad range of cognitive domains. Total administration time was approximately 1 h, and most participants (68%) were administered the CNB in their homes because of the family or subject preference. The PLOT was administered by trained assessors according to standardized instructions and testing conditions, and items were administered in the order that they are listed in Table 1. The full PLOT takes approximately 9 min to administer.

Two hundred and seven subjects did not have valid data for the PLOT and were excluded from analysis, leaving a final *N* of 8,291. The specific criteria for data exclusion were as follows (all based on the full 24-item test):

1. Total task administration time >100 min.

2. Total angle error (across all items) >500°.

3. Total excess mouse clicks (e.g., rotating the line back and forth) >200, or total deficit clicks (clicking too little to even approach a correct answer) >45.

The British sample was administered the PLOT through the Web, and because there were no obvious exclusion criteria given the demographics collected, no one in the British sample older than 18 years of age was excluded from analysis. However, despite preconceived limitations of Internet-based testing, there is evidence that tests administered in person versus online are highly comparable and retain the same psychometric properties (Gosling, Vazire, Srivastava, & John, 2004; Meyerson & Tryon, 2003; Ritter, Lorig, Laurent, & Matthews, 2004).

PLOT item responses were coded in two ways:

1. Dichotomous, such that rotation to a perfectly parallel line set was "correct" (1), and all other responses were "incorrect" (0).

2. Polytomous, such that rotation to a perfectly parallel line set received the maximum score (3), and each mouse click away from perfectly parallel decreased the item score by 1; thus, 3 or more mouse clicks away resulted in no credit (0) for that item.

## Analyses

All analyses described below were performed on both the dichotomous and polytomous response sets. Eigendecompositions were performed on the polychoric correlation matrices to check for sufficient unidimensionality for IRT.

The purpose of the analyses described here was to use simulated CAT to select the best PLOT items for a shortened form. CAT is a method of item administration that updates information about an examinee as he or she responds to items using the response information to determine which item (in a bank of

Table 1
*Full U.S. Sample Factor Analysis and GRM Parameter Estimates for the PLOT24 With Responses Coded As Polytomous and Dichotomous*

| | Item types | | Polytomous | | | | | Dichotomous | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Locations (Difficulties) | | | | | |
| Item | Degrees per click | Line length | Factor loading | Disc | $\delta_1$ | $\delta_2$ | $\delta_3$ | Factor loading | Disc | $\delta_1$ |
| 1 | 9 | L | 0.63 | 0.81 | −2.56 | −1.93 | −0.62 | 0.59 | 0.74 | −0.60 |
| 2 | 9 | L | 0.61 | 0.78 | −1.64 | −1.49 | −0.34 | 0.55 | 0.65 | −0.32 |
| 3 | 9 | M | 0.53 | 0.63 | −2.49 | −1.68 | −0.07 | 0.50 | 0.57 | −0.07 |
| 4 | 9 | M | 0.61 | 0.77 | −2.32 | −1.65 | −0.16 | 0.60 | 0.75 | −0.15 |
| 5 | 6 | M | 0.38 | 0.41 | −1.38 | −0.72 | 0.54 | 0.18 | 0.18 | 0.51 |
| 6 | 6 | L | 0.65 | 0.85 | −1.72 | −1.14 | 0.14 | 0.57 | 0.70 | 0.13 |
| 7 | 6 | M | 0.57 | 0.70 | −1.60 | −0.99 | 0.26 | 0.48 | 0.54 | 0.24 |
| 8 | 6 | L | 0.67 | 0.89 | −1.73 | −1.04 | 0.39 | 0.58 | 0.72 | 0.35 |
| 9 | 3 | L | 0.55 | 0.66 | −0.54 | 0.14 | 1.16 | 0.39 | 0.43 | 1.06 |
| 10 | 3 | M | 0.28 | 0.29 | −0.22 | 0.38 | 1.27 | 0.08 | 0.09 | 1.22 |
| 11 | 3 | M | 0.54 | 0.64 | −0.78 | −0.26 | 0.87 | 0.37 | 0.40 | 0.79 |
| 12 | 3 | L | 0.43 | 0.47 | −0.95 | −0.31 | 0.78 | 0.28 | 0.29 | 0.73 |
| 13 | 3 | L | 0.53 | 0.63 | −0.70 | −0.06 | 1.00 | 0.38 | 0.41 | 0.92 |
| 14 | 9 | L | 0.65 | 0.85 | −2.44 | −1.72 | −0.37 | 0.61 | 0.78 | −0.36 |
| 15 | 9 | M | 0.47 | 0.53 | −2.35 | −1.58 | −0.05 | 0.41 | 0.46 | −0.05 |
| 16 | 3 | L | 0.46 | 0.52 | −0.89 | −0.23 | 0.84 | 0.29 | 0.30 | 0.78 |
| 17 | 6 | L | 0.68 | 0.92 | −1.67 | −0.90 | 0.48 | 0.59 | 0.73 | 0.44 |
| 18 | 9 | L | 0.63 | 0.81 | −1.81 | −1.52 | −0.22 | 0.52 | 0.61 | −0.20 |
| 19 | 6 | L | 0.65 | 0.85 | −1.55 | −0.93 | 0.36 | 0.54 | 0.65 | 0.32 |
| 20 | 6 | M | 0.57 | 0.69 | −1.60 | −0.95 | 0.29 | 0.46 | 0.52 | 0.27 |
| 21 | 9 | M | 0.66 | 0.88 | −2.06 | −1.32 | 0.10 | 0.61 | 0.78 | 0.10 |
| 22 | 3 | M | 0.54 | 0.65 | −0.76 | −0.21 | 0.92 | 0.34 | 0.37 | 0.82 |
| 23 | 3 | M | 0.27 | 0.28 | −0.04 | 0.52 | 1.37 | 0.07 | 0.07 | 1.32 |
| 24 | 6 | M | 0.38 | 0.41 | −1.35 | −0.65 | 0.57 | 0.21 | 0.21 | 0.54 |

*Note.* GRM = graded response model; PLOT = Penn Line Orientation Test; Disc = Discrimination; L = Long; M = Medium. Item response theory (IRT) parameters reported in normal units; to convert to conventional IRT parameters, multiply by 1.702; ratios of first/second eigenvalues for the polytomous and dichotomous models were 5.11 and 3.70, respectively; root mean-square error of approximation for the polytomous and dichotomous models were 0.058 (± 0.001) and 0.051 (± 0.001), respectively; root mean square of the residuals was 0.04 for both the polytomous and dichotomous models.

items) will provide the most information about that examinee. The "most appropriate" item is then administered, and information from the examinee's response is again used to determine the next item to administer, and so on. For example, if an examinee responds correctly to an item of average difficulty, then the adaptive algorithm will temporarily "assume" the examinee is of above-average ability and will select a more difficult item to administer next. If the examinee responds correctly to that second, more difficult item, then the algorithm will update its estimate of the examinee's ability to be even higher and will administer an even more difficult item. This process continues until the examinee responds incorrectly to an item, at which point the algorithm will administer items around that difficulty range until a stopping criterion is met (e.g., the examinee's standard error of measurement reaches some lower threshold). The overall goal of CAT is to avoid administering items that provide very little information about an examinee (for review, see Embretson & Reise, 2000).

Although the above application is focused on the examinee, CAT can also be used to investigate the performance of items within an item bank. For example, if there are some items in the bank that are never administered—either because they are too difficult/easy or because they are not very discriminating—those items might be removed from the bank with no loss in information. Indeed, if the item bank is considered to be the long form of a test, then it might be possible to remove items that are never/rarely administered to create an abbreviated form of that test. Here, we use the long (24-item) version of the PLOT as the item bank and the items' performances in the CAT process to determine whether they are removed to make the short form.

We first fit the Graded Response Model[1] (GRM; Samejima, 1969) to obtain item parameter estimates to later be used in adaptive test simulation (see below). All IRT models were estimated using the irt.fa() command from the psych library (Revelle, 2013) within the R Statistical Package (v3.0.3; R Core Team, 2014).

Estimated item parameters were then input to Firestar (Choi, 2009), an item-response simulation program that allows simulation of CAT sessions, usually in an effort to determine how a particular item bank (and items within that bank) will perform. The user enters the item parameters (in this case, item difficulty and discrimination) for each item and fine tunes certain test specifications (the maximum number of items to administer, which "stopping rule" to use, how to select the next item in the adaptive sequence, how many examinees to simulate, etc.). Firestar then writes an R script to simulate the item responses of *N* examinees and produces several relevant outputs (e.g., which items were administered to each simulated examinee). For the present study, 1,000 examinees were simulated, the maximum number of items to administer was set to 12, and the relevant output was the frequency of each item's administration. These item-administration frequencies were then used to determine which items were eligible for elimination from the final, shortened test form. The above steps were repeated for each (sub-)sample of the PNC: males; females; ages 8–10 years, ages 11–17 years, and ages 18–21 years; and the separate non-U.S. (British) sample.

Shortened, 12-item forms of the PLOT were created for each group, and these groups of 12 items were compared for consistency. A final short form of the PLOT was created based on items that were kept in at least one (sub-)group, resulting in a final 15-item short form. Scores from this short form were then compared across ages and genders to evaluate consistency with previous literature (compared with the full 24-item form).

In addition, Firestar allows users to read in their real data to use for CAT simulation. That is, rather than simulating hypothetical examinees from a normal distribution and then simulating the process of each examinee taking an adaptive version of the test, one can use the real responses given by the actual sample to determine whether an individual would have correctly answered an adaptively administered item. Doing so allows one to avoid the artificial normal distribution used to simulate hypothetical examinees. All simulations described above were performed using such real data simulation, although the results are not shown because they are so similar to the results presented below. Frequencies of item administration changed only minimally, and the end result (i.e., which items were chosen for the final shortened version) did not change at all in any of the samples. The reason for such similarity of results is likely due to the mostly normal distribution of total scores in the real data. If the real distributions of total scores were very skewed or otherwise non-normal, then the results of the hypothetical and real data simulation types could substantially differ.

## Results

Table 1 shows factor analytic[2] and GRM parameter estimates for polytomous and dichotomous responses for the full-length PLOT using the full PNC sample. With only a few exceptions (10 and 23), factor loadings for polytomous items are within the moderate-to-strong range (mean loading = 0.54). Dichotomous items have somewhat weaker loadings ($M$ = 0.43), but as expected, the relative sizes of loadings closely match those of the polytomous items: the correlation between polytomous and dichotomous loadings is .97. Difficulty parameters for the polytomous items tend to be somewhat "easy" (mean difficulty = −0.64), but the upper thresholds ($\delta_3$) are positive overall ($M$ = 0.40), suggesting that the items do provide some information in the upper ability range. Difficulties for dichotomous items cover a range of ability levels, but they provide slightly more information in the upper range (mean difficulty = 0.37). The lower ability coverage of the polytomous items (compared with the dichotomous) is likely because scoring 0 (≥3 clicks off) on a polytomous item is only realistic for very low ability levels (or near complete lack of motivation). Overall, item parameters appear reasonable; thus, the item set is a suitable candidate for the CAT simulation process explored here.

Figure 1 shows the percentage item usage for polytomous items using item parameters estimated in the full sample, with

[1] Technically, the two-parameter logistic model (2PLM; see Embretson & Reise, 2000) was fit to dichotomous responses, but the 2PLM is merely a special case of the GRM (for only two response options).

[2] Factor loadings are reported alongside IRT discrimination parameters because the former are more widely interpretable. Indeed, for the GRM/2PLM used here, there is a direct mathematical translation between factor loadings and discriminations. See Kamata and Bauer (2008) for an explanation of the relationship between factor analytic and IRT parameter estimates.

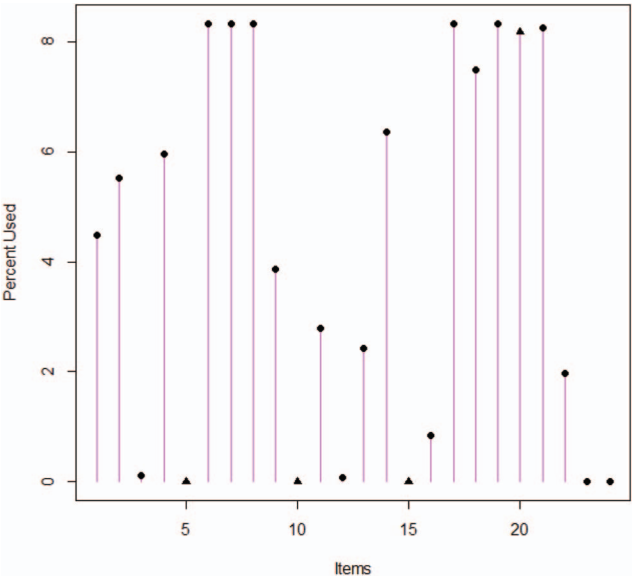

*Figure 1.* Histogram of PLOT24 item usage for adaptive test simulation of 1,000 examinees and maximum item administration of 12 per examinee. See the online article for the color version of this figure.

12 items administered per simulated examinee. The *y*-axis reflects the percentage of total items used; thus, it adds up to 100% for all items. Likewise, item usage of 8.33% (100/12) indicates the item was administered to all 1,000 examinees.

Figure 1 indicates that some items clearly provide more information about examinees; thus, they are more valuable. For example, items 6, 7, 8, 17, and 19 provide so much information that they are always used, regardless of the simulated examinee's ability level (cf. Reise & Henson, 2000). By contrast, items 5, 10, 15, 23, and 24 provide so little information that they are never administered, even when the simulated examinee's ability level is nearly equal to that item's difficulty threshold(s). Such items are obviously candidates for elimination from the battery. Specifically, the results shown in Figure 1 suggest that, if the goal is to create a 12-item short form of the test, then items 3, 5, 9–13, 15, 16, and 22–24 should be removed. Note that such item-usage results were collected for simulations using parameters estimated in each sample (full, male, female, etc.) for a total of seven separate item-elimination recommendations. Here, we chose to be maximally inclusive, eliminating only items that performed well in none of the seven samples' simulations.

Tables 2 and 3 show the final results after the above elimination strategy was implemented in all seven (sub-)samples using the polytomous and dichotomous items. Note that some items (4, 6–8, 14, 17, 19, and 21) had such good parameters that they were selected in all seven samples. By contrast, items 3, 5, 10, 12, 13, 15, 16, 23, and 24 had such poor parameters that they were selected in none of the seven samples. Thus, using a maximally inclusive strategy, all items that were selected by at least one of the samples' simulations were included in the final shortened form. After correcting for item redundancy between forms using Levy's (1967) formula, scores of

Table 2
*PLOT24 Polytomous Item Selection Based on 1,000 Simulated CAT Sessions Using a Normal Distribution of θ, by Sample and Item Type*

| | Item types | | Samples | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | By gender | | By age (Years) | | | | |
| Item | Degrees per click | Line length | Full U.S. sample | Male | Female | 8–10 | 11–17 | 18–21 | British sample | Rem. |
| 1 | 9 | L | S | | | S | S | | S | |
| 2 | 9 | L | S | S | S | S | S | | | |
| 3 | 9 | M | | | | | | | | X |
| 4 | 9 | M | S | S | S | S | S | S | S | |
| 5 | 6 | M | | | | | | | | X |
| 6 | 6 | L | S | S | S | S | S | S | S | |
| 7 | 6 | M | S | S | S | S | S | S | S | |
| 8 | 6 | L | S | S | S | S | S | S | S | |
| 9 | 3 | L | | | S | | | S | | |
| 10 | 3 | M | | | | | | | | X |
| 11 | 3 | M | | S | | | | S | S | |
| 12 | 3 | L | | | | | | | | X |
| 13 | 3 | L | | | | | | | | X |
| 14 | 9 | L | S | S | S | S | S | S | S | |
| 15 | 9 | M | | | | | | | | X |
| 16 | 3 | L | | | | | | | | X |
| 17 | 6 | L | S | S | S | S | S | S | S | |
| 18 | 9 | L | S | S | S | S | S | S | | |
| 19 | 6 | L | S | S | S | S | S | S | S | |
| 20 | 6 | M | S | S | S | S | S | | S | |
| 21 | 9 | M | S | S | S | S | S | S | S | |
| 22 | 3 | M | | | | | | S | S | |
| 23 | 3 | M | | | | | | | | X |
| 24 | 6 | M | | | | | | | | X |

*Note.* PLOT = Penn Line Orientation Test; CAT = computerized adaptive testing; L = Long; M = Medium; S = selected; Rem. = removed from final PLOT15 battery.

Table 3
*PLOT24 Dichotomous Item Selection Based on 1,000 Simulated CAT Sessions Using a Normal Distribution of* θ*, by Sample Type*

| | Item types | | Samples | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | By gender | | By age (Years) | | | | |
| Item | Degrees per click | Line length | Full U.S. sample | Male | Female | 8–10 | 11–17 | 18–21 | British sample | Rem. |
| 1 | 9 | L | S | S | S | S | S | S | S | |
| 2 | 9 | L | S | S | S | S | S | S | | |
| 3 | 9 | M | S | S | S | | S | S | S | |
| 4 | 9 | M | S | S | S | S | S | S | S | |
| 5 | 6 | M | | | | | | | | X |
| 6 | 6 | L | S | S | S | S | S | S | S | |
| 7 | 6 | M | S | S | S | S | S | S | | |
| 8 | 6 | L | S | S | S | S | S | S | S | |
| 9 | 3 | L | | | | | | | | X |
| 10 | 3 | M | | | | | | | | X |
| 11 | 3 | M | | | | | | | | X |
| 12 | 3 | L | | | | | | | | X |
| 13 | 3 | L | | | | | | | | X |
| 14 | 9 | L | S | S | S | S | S | S | S | |
| 15 | 9 | M | | | | | | | S | |
| 16 | 3 | L | | | | | | | | X |
| 17 | 6 | L | S | S | S | S | S | S | S | |
| 18 | 9 | L | S | S | S | S | S | S | S | |
| 19 | 6 | L | S | S | S | S | S | S | S | |
| 20 | 6 | M | | | | S | | | S | |
| 21 | 9 | M | S | S | S | S | S | S | S | |
| 22 | 3 | M | | | | | | | | X |
| 23 | 3 | M | | | | | | | | X |
| 24 | 6 | M | | | | | | | | X |

*Note.* PLOT = Penn Line Orientation Test; CAT = computerized adaptive testing; L = Long; M = Medium; S = selected; Rem. = removed from final PLOT15 battery.

the shortened form correlated .90 with scores from the full form. Cronbach's α values for the full and shortened form were 0.92 and 0.91, respectively, when polytomous scoring was used (see Table 2). When dichotomous scoring was used, Cronbach's α values for the full and shortened forms were 0.85 and 0.86, respectively. From here on, we report results based only on the polytomous item scoring.

One useful way to show the similarity between a full-length test and its shortened form is to compare their test information curves (TICs). Figure 2 shows the TICs for the PLOT24 and PLOT15, with reliability on the dual *y*-axis. Of note, both curves have nearly identical shapes, with maximum reliability of 0.87 and 0.84 for the long and short versions, respectively. In addition, the location of maximum information for both tests is approximately −1.0, suggesting that they are optimal for individuals of slightly below-average ability. This is consistent with the negative skew of the sum scores (not shown).

Table 4 shows the sex and age effects for the full-length and shortened versions' polytomous scores. As expected for this spatial task (Gur et al., 2010, 2012), males outperform females by 3.7% using the full-length scores and 3.1% using the shortened scores, with both effects significant at the $p < .001$ level. Correlation with age, which is expected in this developmental (PNC) sample, is also nearly equal for both scores (.402 for full-length scores and .408 for shortened scores), with both values again significant at the $p < .001$ level.

## Discussion

In this study, we used a novel IRT-based CAT simulation technique to develop an abbreviated version of the PLOT, which is a computerized assessment of visuospatial perception originally targeted by Benton's classic judgment of line orientation test (Benton et al., 1978). Although the PLOT has been shown to possess adequate psychometric characteristics and to evidence validity (see below), it is nonetheless possible that some items on the test are not as useful as others in assessing the underlying latent trait. IRT offers the ability to analyze which individual test items "work best" in assessing an underlying trait or ability by incorporating information about discriminative ability and difficulty of each item. IRT provides many advantages in attempting to design shortened versions of previously validated instruments. A key advantage of IRT is that it accounts for difficulty and discrimination whereas a conventional (correlational) approach accounts only for the latter. For example, if one were to assess the quality of items only on the basis of their correlation with total score, he or she would run the risk of choosing items all within the same difficulty range. By contrast, the IRT-based CAT simulation approach used here attempts to balance the importance of having highly discriminating items with the importance of having items that cover a wide range of difficulty levels. This balance is accomplished by simulating examinees from a normal ability distribution, such that highly discriminating items are selected only if they fall within the reasonable range of

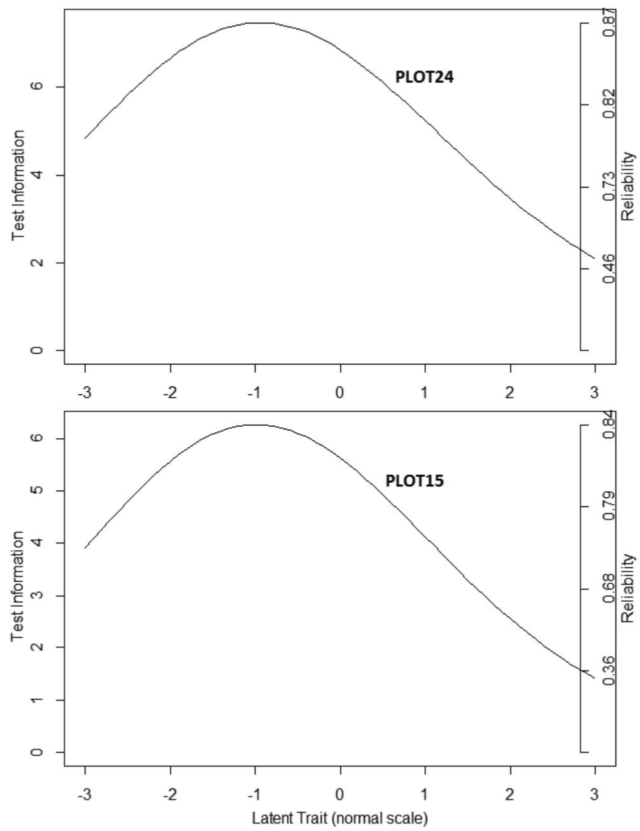

*Figure 2.* Test information curves for the Penn Line Orientation Test (PLOT24) and its shortened version, the PLOT15.

difficulty. Likewise, an item with only moderate discriminatory ability might nonetheless be chosen if its difficulty parameter is ideally placed on the ability continuum (e.g., an item with exactly average difficulty).

Initial factor analytic approaches indicated that for both the polytomous and dichotomous scoring, the PLOT had overall strong factor loadings and ratios of first/second eigenvalues >3.0, indicating that the common variance among the PLOT items is explained mostly by a single factor. The PLOT's unidimensional structure is important for two reasons. First, it indicates that the PLOT item set is appropriate for our CAT simulation approach. Second, the fact that the measure appears unidimensional diminishes concerns that short forms of an instrument are often inferior in examining factors within the larger scale (e.g., Smith, McCarthy, & Anderson, 2000).

Follow-up analyses with the IRT-based CAT model determined that there was a consistent set of five items that were maximally useful across all ability levels. Likewise, five items were not useful across any ability levels, even when the examinee ability was almost equivalent to the item's difficulty threshold. Overall, IRT models showed that item discrimination was fairly consistent across a relatively wide age range, across sex, and across both a U.S.-based sample of youth and a British sample of adults collected via divergent methods. Overall, our models were relatively consistent in identifying which items were useful and which items would not provide important information about both the examinee and the underlying construct.

Our results point to the potential utility of this novel approach for identifying items that can be eliminated from a set of assessment items. As mentioned above, one strength of the approach is that it simultaneously emphasizes item discrimination and difficulty. In addition, this method allows a researcher to choose an expected distribution of trait levels, ensuring that items with the appropriate difficulty levels are given a slight advantage in the item selection process. Here, the simulated trait distribution was normal (i.e., it was assumed that average examinees were common, and examinees in the "tails" of the distribution were rare); however, the Firestar program allows simulation of any distribution type. Thus, if a researcher knew that a particular trait distribution was highly skewed—or even multimodal—he or she could simulate examinees from that particular distribution type.

Using IRT simulation techniques and taking the maximally inclusive approach of only removing the items that were the least informative across all of the samples (i.e., not chosen as performing well in any sample), we were able to shorten the PLOT from 24 to 15 items, which reduces administration time from approximately 9 min to approximately 5 min. This reduced administration time increases the measure's feasibility in deployment for large-scale studies, especially those in which visuospatial processing ability may not be a primary outcome. Of note, as shown in Tables 2 and 3, the items chosen for the PLOT15 cover a relatively well-distributed range of length of lines, although the IRT models were not specifically designed to accomplish this distribution. However, these models did eliminate a greater number of the 3° per click items, which may indicate that these items do not effectively assess the construct

Table 4
*Means (and Percentages) Correct and Correlations With Age for the Standard and Short PLOT Showing Sex and Age Effects for the Full-Length (24-Item) and Shortened (15-Item) Versions of the PLOT*

| | Sex differences | | | | | Age trends | | |
|---|---|---|---|---|---|---|---|---|
| | Means | | | | | | | |
| Score | Male | Female | Diff (M – F) | $t$ | $p$ | Correlation with age | $t$ | $p$ |
| Full-length | 48.12 (66.8%) | 45.46 (63.1%) | 2.66 (3.7%) | 10.3 | <.001 | 0.402 | 40.0 | <.001 |
| Shortened | 32.07 (71.3%) | 30.70 (68.2%) | 1.37 (3.1%) | 7.4 | <.001 | 0.408 | 40.7 | <.001 |

*Note.* PLOT = Penn Line Orientation Test; Diff = difference; M = Male; F = Female; Corr = Correlation. Means based on polytomous scores, such that the highest possible score on the 24-item test was 72 (score of 3 on all items) and for the 15-item test was 45.

underlying performance on the PLOT compared with the easier 6° and 9° per click items, assuming a normal trait distribution. Such items could be used in discriminating among high-performing samples.

It is important that an abbreviated version of a neurocognitive instrument not only correlates with the total score of the full version but (also) provides similar discrimination of clinical groups or individual differences. Therefore, we investigated whether the abbreviated version of the PLOT displayed similar sensitivity to sex and age differences when compared with the full version of the measure. We found that the PLOT15 showed the same magnitude of sex differences and the same correlation with age compared with the full, 24-item version in the large PNC sample. This is consistent with previous findings (e.g., Gur et al., 2012).

One topic that has not been discussed thus far is the validity of the PLOT. The American Educational Research Association has developed standards for how the validity of a test should be evaluated, and they suggest that evidence for validity should fall into one of five "types" (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational & Psychological Testing, 1999). One type is the relationship of the test's scores to external variables, such as neurological phenomena and/or scores on similar tests. This type of evidence for the PLOT's validity is abundant: Roalf et al. (2014) demonstrated that performance on the PLOT is associated with activity in hypothesized brain areas (also see Satterthwaite et al., 2013), and Moore and colleagues (2014) demonstrated that the PLOT correlated more highly with other tests within its neurocognitive domain (complex reasoning) than with tests designed to measure other domains (e.g., memory or social cognition). Another type of evidence, sometimes called "structural validity," relates to whether the test components (individual items, in this case) relate to each other in ways consistent with the theory used to construct it. Because the PLOT is designed to measure only one construct, evidence for structural validity would consist of demonstrating that the measure is unidimensional. The ratio of first to second eigenvalues and the moderate fit of the unidimensional model (see Table 1) provide some such evidence. Finally, a third type of evidence, sometimes called "face validity," relates to the consistency of test content (e.g., item wording) with theory and common sense. This type of evidence is less intuitive for neurocognitive tests, in which there often is no "item content" as such, but it is worth noting that the design of the PLOT is based on a well-established, decades-old paradigm for assessing visuospatial ability (Benton et al., 1978), which has itself accumulated evidence of validity such as theoretically consistent correlations of scores with disease (e.g., Montse, Pere, Carme, Francesc, & Eduardo, 2001), cerebral blood flow (Gur et al., 1982, 1994, 2000; Hannay et al., 1987), structural neuroanatomy (Tranel et al., 2009), and brain lesions (Trahan, 1998).

There are limitations to this study. Although we describe a simulation model in which certain items are not administered because of their failure to provide useful information about the examinee, the samples described were administered the full versions of these tests. Whether there is an effect of not administering the items that were removed in our simulations can only be addressed by administering both versions of the test to the same participants. Relatedly, the reliability of scores and the validity of test score interpretation of the 15-item version of the PLOT will need to be investigated further because one cannot assume that the PLOT15 inherently possesses the same psychometric characteristics as the 24-item version (Smith et al., 2000).

Despite these limitations, we were able to develop an abbreviated version of the PLOT that maximized the utility of items across two large, independent samples by taking into account both the discriminability and difficulty of each item. Brief but valid assessments of neurocognitive abilities are increasingly needed in large-scale clinical, treatment, and genomic studies, and the abbreviated version of the PLOT developed here would be appropriate for investigations in which visuospatial processing may not be the primary focus of study but its adequate assessment is nonetheless desired. It is important to note that the test is freely available online for qualified investigators who want to use it in research with institutional review board oversight (http://www.med.upenn.edu/bbl/).

## References

Aliyu, M. H., Calkins, M. E., Swanson, C. L., Jr., Lyons, P. D., Savage, R. M., May, R., . . . Allen, T. B., & the PAARTNERS Study Group. (2006). Project among African-Americans to explore risks for schizophrenia (PAARTNERS): Recruitment and assessment methods. *Schizophrenia Research, 87*(1–3): 32–44. http://dx.doi.org/10.1016/j.schres.2006.06.027

Almasy, L., Gur, R. C., Haack, K., Cole, S. A., Calkins, M. E., Peralta, J. M., . . . Gur, R. E. (2008). A genome screen for quantitative trait loci influencing schizophrenia and neurocognitive phenotypes. *The American Journal of Psychiatry, 165,* 1185–1192.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Benton, A. L., Varney, N. R., & Hamsher, K. D. (1978). Visuospatial judgment. A clinical test. *Archives of Neurology, 35,* 364–367. http://dx.doi.org/10.1001/archneur.1978.00500300038006

Calkins, M. E., Moore, T. M., Merikangas, K. R., Burstein, M., Satterthwaite, T. D., Bilker, W. B., . . . Gur, R. E. (2014). The psychosis spectrum in a young U.S. community sample: Findings from the Philadelphia Neurodevelopmental Cohort. *World Psychiatry; Official Journal of the World Psychiatric Association (WPA), 13,* 296–305. http://dx.doi.org/10.1002/wps.20152

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement, 33,* 644–645. http://dx.doi.org/10.1177/0146621608329892

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59,* 93–104. http://dx.doi.org/10.1037/0003-066X.59.2.93

Grant, P. M., Huh, G. A., Perivoliotis, D., Stolar, N. M., & Beck, A. T. (2012). Randomized trial to evaluate the efficacy of cognitive therapy for low-functioning patients with schizophrenia. *Archives of General Psychiatry, 69,* 121–127. http://dx.doi.org/10.1001/archgenpsychiatry.2011.129

Greenwood, T. A., Braff, D. L., Light, G. A., Cadenhead, K. S., Calkins, M. E., Dobie, D. J., . . . Schork, N. J. (2007). Initial heritability analyses of endophenotypic measures for schizophrenia: The consortium on the genetics of schizophrenia. *Archives of General Psychiatry, 64,* 1242–1250. http://dx.doi.org/10.1001/archpsyc.64.11.1242

Gur, R. C., Alsop, D., Glahn, D., Petty, R., Swanson, C. L., Maldjian, J. A., . . . Gur, R. E. (2000). An fMRI study of sex differences in regional activation to a verbal and a spatial task. *Brain and Language, 74,* 157–170. http://dx.doi.org/10.1006/brln.2000.2325

Gur, R. E., Calkins, M. E., Gur, R. C., Horan, W. P., Nuechterlein, K. H., Seidman, L. J., & Stone, W. S. (2007). The Consortium on the Genetics of Schizophrenia: Neurocognitive endophenotypes. *Schizophrenia Bulletin, 33,* 49–68. http://dx.doi.org/10.1093/schbul/sbl055

Gur, R. C., Erwin, R. J., & Gur, R. E. (1992). Neurobehavioral probes for physiologic neuroimaging studies. *Archives of General Psychiatry, 49,* 409–414. http://dx.doi.org/10.1001/archpsyc.1992.01820050073013

Gur, R. C., Gur, R. E., Obrist, W. D., Hungerbuhler, J. P., Younkin, D., Rosen, A. D., . . . Reivich, M. (1982). Sex and handedness differences in cerebral blood flow during rest and cognitive activity. *Science, 217,* 659–661. http://dx.doi.org/10.1126/science.7089587

Gur, R. E., Nimgaonkar, V. L., Almasy, L., Calkins, M. E., Ragland, J. D., Pogue-Geile, M. F., . . . Gur, R. C. (2007). Neurocognitive endophenotypes in a multiplex multigenerational family study of schizophrenia. *The American Journal of Psychiatry, 164,* 813–819. http://dx.doi.org/10.1176/ajp.2007.164.5.813

Gur, R. C., Ragland, J. D., Moberg, P. J., Turner, T. H., Bilker, W. B., Kohler, C., . . . Gur, R. E. (2001). Computerized neurocognitive scanning: I. Methodology and validation in healthy people. *Neuropsychopharmacology, 25,* 766–776. http://dx.doi.org/10.1016/S0893-133X(01)00278-0

Gur, R. C., Ragland, J. D., Resnick, S. M., Skolnick, B. E., Jaggi, J., Muenz, L., & Gur, R. E. (1994). Lateralized increases in cerebral blood flow during performance of verbal and spatial tasks: Relationship with performance level. *Brain and Cognition, 24,* 244–258. http://dx.doi.org/10.1006/brcg.1994.1013

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. http://dx.doi.org/10.1037/a0026712

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. http://dx.doi.org/10.1016/j.jneumeth.2009.11.017

Hannay, H. J., Falgout, J. C., Leli, D. A., Katholi, C. R., Halsey, J. H., Jr., & Wills, E. L. (1987). Focal right temporo-occipital blood flow changes associated with judgment of line orientation. *Neuropsychologia, 25,* 755–763. http://dx.doi.org/10.1016/0028-3932(87)90113-8

Iannacone, S., Leary, M., Esposito, E. C., Ruparel, K., Savitt, A., Mott, A., . . . Abella, B. S. (2014). Feasibility of cognitive functional assessment in cardiac arrest survivors using an abbreviated laptop-based neurocognitive battery. *Therapeutic Hypothermia and Temperature Management, 4,* 131–136. doi:10.1089/ther.2014.0007

Insel, T. R., & Cuthbert, B. N. (2009). Endophenotypes: Bridging genomic complexity and disorder heterogeneity. *Biological Psychiatry, 66,* 988–989. http://dx.doi.org/10.1016/j.biopsych.2009.10.008

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15,* 136–153. http://dx.doi.org/10.1080/10705510701758406

Levy, P. (1967). The correction for spurious correlation in the evaluation of short-form tests. *Journal of Clinical Psychology, 23,* 84–86. http://dx.doi.org/10.1002/1097-4679(196701)23:1<84::AID-JCLP2270230123>3.0.CO;2-2

Merikangas, K. R., Calkins, M. E., Burstein, M., He, J.-P., Chiavacci, R., Lateef, T., . . . Gur, R. E. (in press). Comorbidity of physical and mental disorders in the Neurodevelopmental Genomics Cohort Study. *Pediatrics*.

Meyerson, P., & Tryon, W. W. (2003). Validating internet research: A test of the psychometric equivalence of internet and in-person samples. *Behavior Research Methods, Instruments & Computers, 35,* 614–620. http://dx.doi.org/10.3758/BF03195541

Montse, A., Pere, V., Carme, J., Francesc, V., & Eduardo, T. (2001). Visuospatial deficits in Parkinson's disease assessed by judgment of line orientation test: Error analyses and practice effects. *Journal of Clinical and Experimental Neuropsychology, 23,* 592–598. http://dx.doi.org/10.1076/jcen.23.5.592.1248

Moore, T. M., Reise, S. P., Gur, R. E., Hakonarson, H., & Gur, R. C. (2014). Psychometric properties of the Penn Computerized Neurocognitive Battery. [Advance online publication]. *Neuropsychology*. http://dx.doi.org/10.1037/neu0000093

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20,* 33–65. http://dx.doi.org/10.1016/j.acn.2004.02.005

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7,* 347–364. http://dx.doi.org/10.1177/107319110000700404

Revelle, W. (2013). *Psych: Procedures for personality and psychological research* (Version 1.4.2). Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych

Ritter, P., Lorig, K., Laurent, D., & Matthews, K. (2004). Internet versus mailed questionnaires: A randomized comparison. *Journal of Medical Internet Research, 6,* e29. http://dx.doi.org/10.2196/jmir.6.3.e29

Roalf, D. R., Ruparel, K., Gur, R. E., Bilker, W., Gerraty, R., Elliott, M. A., . . . Gur, R. C. (2014). Neuroimaging predictors of cognitive performance across a standardized neurocognitive battery. *Neuropsychology, 28,* 161–176. http://dx.doi.org/10.1037/neu0000011

Roalf, D. R., Ruparel, K., Verma, R., Elliott, M. A., Gur, R. E., & Gur, R. C. (2013). White matter organization and neurocognitive performance variability in schizophrenia. *Schizophrenia Research, 143,* 172–178. http://dx.doi.org/10.1016/j.schres.2012.10.014

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34,* 100.

Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M. E., . . . Gur, R. E. (2014). Neuroimaging of the Philadelphia neurodevelopmental cohort. *NeuroImage, 86,* 544–553. http://dx.doi.org/10.1016/j.neuroimage.2013.07.064

Satterthwaite, T. D., Wolf, D. H., Erus, G., Ruparel, K., Elliott, M. A., Gennatas, E. D., . . . Gur, R. E. (2013). Functional maturation of the executive system during adolescence. *The Journal of Neuroscience, 33,* 16249–16261. http://dx.doi.org/10.1523/JNEUROSCI.2345-13.2013

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429–438). New York, NY: Academic Press. http://dx.doi.org/10.1016/B0-12-369398-5/00444-8

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12,* 102–111. http://dx.doi.org/10.1037/1040-3590.12.1.102

Spencer, R. J., Wendell, C. R., Giggey, P. P., Seliger, S. L., Katzel, L. I., & Waldstein, S. R. (2013). Judgment of Line Orientation: An examination of eight short forms. *Journal of Clinical and Experimental Neuro-*

*psychology, 35,* 160–166. http://dx.doi.org/10.1080/13803395.2012.760535

Thomas, P., Bhatia, T., Gauba, D., Wood, J., Long, C., Prasad, K., . . . Deshpande, S. N. (2013). Exposure to herpes simplex virus, type 1 and reduced cognitive function. *Journal of Psychiatric Research, 47,* 1680–1685. http://dx.doi.org/10.1016/j.jpsychires.2013.07.010

Trahan, D. E. (1998). Judgment of line orientation in patients with unilateral cerebrovascular lesions. *Assessment, 5,* 227–235. http://dx.doi.org/10.1177/107319119800500303

Tranel, D., Vianna, E., Manzel, K., Damasio, H., & Grabowski, T. (2009). Neuroanatomical correlates of the Benton Facial Recognition Test and Judgment of Line Orientation Test. *Journal of Clinical and Experimental Neuropsychology, 31,* 219–233. http://dx.doi.org/10.1080/13803390802317542

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., . . . Yacoub, E., & the WU-Minn HCP Consortium. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage, 62,* 2222–2231. http://dx.doi.org/10.1016/j.neuroimage.2012.02.018

Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361–375. http://dx.doi.org/10.1111/j.1745-3984.1984.tb01040.x

Received November 6, 2014
Revision received January 20, 2015
Accepted January 20, 2015 ■