HHS Public Access

Author manuscript

Psychol Assess. Author manuscript; available in PMC 2016 September 01.

Published in final edited form as:

Psychol Assess. 2015 September ; 27(3): 955–964. doi: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[a], **J. Cobb Scott**[a,b], **Steven P. Reise**[c], **Allison M. Port**[a], **Chad T. Jackson**[a], **Kosha Ruparel**[a], **Adam P. Savitt**[a], **Raquel E. Gur**[a], and **Ruben C. Gur**[a,b]

[a] Department of Psychiatry, Brain Behavior Laboratory, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

[b] VISN4 Mental Illness Research, Education, and Clinical Center at the Philadelphia VA Medical Center, Philadelphia, PA, 19104, USA

[c] Department of Psychology, University of California, Los Angeles, CA, USA.

Abstract

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 to 21) 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 television 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 one 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

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

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; R. C. 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).

Due to 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 (Penn 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 (Roalf et al., 2014) and with established psychometric properties (Moore, Reise, Gur, Hakonarson, & Gur, 2014). Although the tests that compose the Penn 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., R. C. Gur et al., 2000; Roalf et al., 2013, 2014; Satterthwaite et al., 2013), making them particularly useful as biomarkers of brain dysfunction (R. C. 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; R. C. Gur et al., 2001, 2012; R. E. Gur, Calkins, et al., 2007; R. E. Gur, Nimgaonkar, et al., 2007; Thomas et al., 2013).

The Penn Line Orientation Test (PLOT) is included in the Penn 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 9^0 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 (R. C. Gur et al., 2001, 2010; Moore, Reise, Gur, Hakonarson, & Gur, 2014) and has been used in large-scale genomic studies to examine associations with psychiatric disorders and brain structure and function (R. C. 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. Additionally, while 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, Satterthwaite et al., 2014, Merikangas et al., in press). 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 television documentary, and we examine the consistency of this abbreviated form with previous literature in detecting age and sex differences in visuospatial processing.

Methods

Participants and Settings

This study includes two independent samples. The first sample comprises 8,498 youths ages 8 to 21 (51% Female; 57% Caucasian; mean age = 13.4) who were administered a battery of neurocognitive tests as part of their participation in the NIMH-funded Philadelphia Neurodevelopmental Cohort (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: (1) being able to provide informed consent; (2) proficiency in English; and (3) 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 in order to recruit a representative sample of youth from the greater Philadelphia area (see Calkins et al., 2014). Participants and their guardians (for participants under 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 television viewers (77% female; 91% Caucasian; mean age = 34.1) in the UK who were invited to take the test over the web after a television documentary on sex differences was aired on the BBC. Participants were shown a brief textual explanation of the task itself as well as 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 decline to participate, in which case they were routed back to the television program's website. Participants under the age of 18 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 (R. C. Gur et al., 2012; Moore, Reise, Gur, Hakonarson, & Gur, 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 hour, 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 they are listed in Table 1. The full PLOT takes approximately nine minutes 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 8291. The specific criteria for data exclusion were as follows (all based on the full 24-item test):

a) Total task administration time > 100 minutes.

b) Total angle error (across all items) > 500 degrees.

c) 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 were administered the PLOT through the web, and because there were no obvious exclusion criteria given the demographics collected, no one in the British sample over the age of 18 was excluded from analysis. Despite preconceived limitations of internet-based testing, however, 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 item response theory (IRT).

The purpose of the analyses described here was to use simulated computerized adaptive testing (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/she responds to items, using the response information to determine which item (in a bank of 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, 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, 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).

While 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 computerized adaptive testing 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, 1000 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

[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).

(sub-)sample of the PNC: males, females, ages 8-10, ages 11-17, ages 18-21; and the separate non-US (British) sample.

Shortened, 12-item forms of the PLOT were created for each group, and these groups of twelve 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 to the full 24-item form).

Additionally, 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 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 answered an adaptively administered item correctly. 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, though 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, the results of the hypothetical and real data simulation types could differ substantially.

Results

Table 1 shows factor analytic[2] and GRM parameter estimates for polytomous and dichotomous responses for the full-length VSPLOT 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 (mean = 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 0.97. Difficulty parameters for the polytomous items tend to be somewhat “easy” (mean difficulty = −0.64), but the upper thresholds (δ_3) are positive overall (mean = 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 provide slightly more information in the upper range (mean difficulty = 0.37). The lower ability coverage of the polytomous items (compared to the dichotomous) is likely due to the fact that scoring 0 (3+ clicks off) on a polytomous item is realistic only for very low ability levels (or near complete lack of motivation). Overall, item parameters appear reasonable, and the item set is thus a suitable candidate for the CAT simulation process explored here.

Figure 1 shows the percent item usage for polytomous items using item parameters estimated in the full sample, with 12 items administered per simulated examinee. The y-axis

[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.

reflects the percent of total items used, and thus adds up to 100% for all items. Likewise, item usage of 8.33% (100/12) indicates the item was administered to all 1000 examinees. Figure 1 indicates that some items clearly provide more information about examinees, and are thus 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, 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 was included in the final shortened form. After correcting for item redundancy between forms using Levy's (1967) formula, scores of the shortened form correlated 0.90 with scores from the full form. Cronbach's α for the full and shortened form were 0.92 and 0.91, respectively when polytomous scoring was used (Table 2). When dichotomous scoring was used, Cronbach's α 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. Notably, both curves have nearly identical shapes, with maximum reliability of 0.87 and 0.84 for the long and short versions, respectively. Also, the location of maximum information for both tests is around -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 $p < 0.001$ level. Correlation with age, which is expected in this developmental (PNC) sample, is also nearly equal for both scores: 0.402 for full-length scores and 0.408 for shortened scores, with both values again significant at the $p < 0.001$ level.

Discussion

In this study, we used a novel IRT-based CAT simulation technique to develop an abbreviated version of the Penn Line Orientation Test, which is a computerized assessment of visuospatial perception originally targeted by Benton's classic judgment of line orientation test (Benton, Varney, & Hamsher, 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 both difficulty and discrimination, whereas a conventional (correlational) approach accounts only for the latter. If one were to assess the quality of items based only on their correlation with total score, for example, he/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 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 $1^{st}/2^{nd}$ 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. Similarly, 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 US-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 emphasizes both item discrimination and difficulty simultaneously.

Additionally, 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/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 nine minutes to approximately five minutes. 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. Notably, as shown in Tables 2 and 3, the items chosen for the PLOT-15 cover a relatively well-distributed range of length of lines even though the IRT models were not specifically designed to accomplish this distribution. However, these models did eliminate a greater number of the 3-degree per click items, which may indicate that these items do not effectively assess the construct underlying performance on the PLOT compared to the easier 6- and 9-degree 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. We therefore investigated whether the abbreviated version of the PLOT displayed similar sensitivity to sex and age differences when compared to the full version of the measure. We found that the PLOT-15 showed the same magnitude of sex differences and the same correlation with age compared to 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" (AERA et al., 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, Reise, Gur, Hakonarson, & Gur (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 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, where 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 visuo-spatial ability (Benton, Varney, & Hamsher, 1978), which has itself accumulated evidence of validity such as theoretically-consistent correlations of scores with disease (e.g. Montse et al., 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 impact 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 validity of test score interpretation of the 15-item version of the PLOT will need to be investigated further, as one cannot assume that the PLOT-15 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. Importantly, 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/).

Acknowledgments

This work was supported by NIMH grants MH089983, MH019112, MH096891

References

Aliyu MH, Calkins ME, Swanson CL, Lyons PD, Savage RM, May R, PAARTNERS Study Group. Project among African-Americans to explore risks for schizophrenia (PAARTNERS): recruitment and assessment methods. Schizophrenia Research. 2006; 87(1-3):32–44. doi:10.1016/j.schres.2006.06.027. [PubMed: 16887335]

Almasy L, Gur RC, Haack K, Cole SA, Calkins ME, Peralta JM, Gur RE. A genome screen for quantitative trait loci influencing schizophrenia and neurocognitive phenotypes. The American Journal of Psychiatry. 2008; 165(9):1185–1192. doi:10.1176/appi.ajp.2008.07121869. [PubMed: 18628350]

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational & Psychological Testing (US). Standards for educational and psychological testing. Amer Educational Research Assn.; 1999.

Benton AL, Varney NR, Hamsher KD. Visuospatial judgment. A clinical test. Archives of Neurology. 1978; 35(6):364–367. [PubMed: 655909]

Calkins ME, Moore TM, Merikangas KR, Burstein M, Satterthwaite TD, Bilker WB, Gur RE. The psychosis spectrum in a young US community sample: findings from the Philadelphia Neurodevelopmental Cohort. World Psychiatry. 2014; 13(3):296–305. [PubMed: 25273303]

Choi SW. Firestar: Computerized Adaptive Testing Simulation Program for Polytomous Item Response Theory Models. Applied Psychological Measurement. 2009; 33(8):644–645. doi: 10.1177/0146621608329892.

Embretson, SE.; Reise, SP. Item response theory for psychologists. Lawrence Erlbaum; Mahwah, NJ: 2000.

Gosling SD, Vazire S, Srivastava S, John OP. Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. American Psychologist. 2004; 59(2): 93. [PubMed: 14992636]

Grant PM, Huh GA, Perivoliotis D, Stolar NM, Beck AT. Randomized trial to evaluate the efficacy of cognitive therapy for low-functioning patients with schizophrenia. Archives of General Psychiatry. 2012; 69(2):121–127. doi:10.1001/archgenpsychiatry.2011.129. [PubMed: 21969420]

Greenwood TA, Braff DL, Light GA, Cadenhead KS, Calkins ME, Dobie DJ, Schork NJ. Initial heritability analyses of endophenotypic measures for schizophrenia: the consortium on the genetics of schizophrenia. Archives of General Psychiatry. 2007; 64(11):1242–1250. doi:10.1001/archpsyc.64.11.1242. [PubMed: 17984393]

Gur RC, Gur RE, Obrist WD, Hungerbuhler JP, Younkin D, Rosen AD, Reivich M. Sex and handedness differences in cerebral blood flow during rest and cognitive activity. Science. 1982; 217(4560):659–661. [PubMed: 7089587]

Gur RC, Ragland JD, Resnick SM, Skolnick BE, Jaggi J, Muenz L, Gur RE. Lateralized increases in cerebral blood flow during performance of verbal and spatial tasks: Relationship with performance level. Brain and Cognition. 1994; 24(2):244–258. [PubMed: 8185896]

Gur RC, Alsop D, Glahn D, Petty R, Swanson CL, Maldjian JA, Gur RE. An fMRI study of sex differences in regional activation to a verbal and a spatial task. Brain and Language. 2000; 74(2): 157–170. doi:10.1006/brln.2000.2325. [PubMed: 10950912]

Gur RC, Erwin RJ, Gur RE. Neurobehavioral probes for physiologic neuroimaging studies. Archives of General Psychiatry. 1992; 49:409–414. [PubMed: 1586277]

Gur RC, Gur RE, Obrist WD, Hungerbuhler JP, Younkin D, Rosen AD, Reivich M. Sex and handedness differences in cerebral blood flow during rest and cognitive activity. Science. 1982; 217(4560):659–661. [PubMed: 7089587]

Gur RC, Ragland JD, Moberg PJ, Turner TH, Bilker WB, Kohler C, Gur RE. Computerized neurocognitive scanning: I. Methodology and validation in healthy people. Neuropsychopharmacology. 2001; 25(5):766–776. doi:10.1016/S0893-133X(01)00278-0. [PubMed: 11682260]

Gur RC, Richard J, Calkins ME, Chiavacci R, Hansen JA, Bilker WB, Gur RE. Age group and sex differences in performance on a computerized neurocognitive battery in children age 8-21. Neuropsychology. 2012; 26(2):251–265. doi:10.1037/a0026712. [PubMed: 22251308]

Gur RC, Richard J, Hughett P, Calkins ME, Macy L, Bilker WB, Gur RE. A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: standardization and initial construct validation. Journal of Neuroscience Methods. 2010; 187(2):254–262. doi: 10.1016/j.jneumeth.2009.11.017. [PubMed: 19945485]

Gur RE, Calkins ME, Gur RC, Horan WP, Nuechterlein KH, Seidman LJ, Stone WS. The Consortium on the Genetics of Schizophrenia: neurocognitive endophenotypes. Schizophrenia Bulletin. 2007; 33(1):49–68. doi:10.1093/schbul/sbl055. [PubMed: 17101692]

Gur RE, Nimgaonkar VL, Almasy L, Calkins ME, Ragland JD, Pogue-Geile MF, Gur RC. Neurocognitive endophenotypes in a multiplex multigenerational family study of schizophrenia. The American Journal of Psychiatry. 2007; 164(5):813–819. doi:10.1176/appi.ajp.164.5.813. [PubMed: 17475741]

Hannay HJ, Falgout JC, Leli DA, Katholi CR, Halsey JH, Wills EL. Focal right temporo-occipital blood flow changes associated with judgment of line orientation. Neuropsychologia. 1987; 25(5): 755–763. [PubMed: 3431672]

Iannacone S, Leary M, Esposito EC, Ruparel K, Savitt A, Mott A, Abella BS. Feasibility of Cognitive Functional Assessment in Cardiac Arrest Survivors Using an Abbreviated Laptop-Based Neurocognitive Battery. Therapeutic Hypothermia and Temperature Management. 2014 doi: 10.1089/ther.2014.0007.

Insel TR, Cuthbert BN. Endophenotypes: Bridging genomic complexity and disorder heterogeneity. Biological Psychiatry. 2009; 66(11):988–989. [PubMed: 19900610]

Kamata A, Bauer DJ. A Note on the Relation Between Factor Analytic and Item Response Theory Models. Structural Equation Modeling: A Multidisciplinary Journal. 2008; 15(1):136–153. doi: 10.1080/10705510701758406.

Levy P. The correction for spurious correlation in the evaluation of short-form tests. Journal of Clinical Psychology. 1967; 23(1):84–86. [PubMed: 6031038]

Merikangas KR, Calkins ME, Burstein M, He J-P, Chiavacci R, Lateef T, Ruparel K, Gur RC, Lehner T, Hakonarson K, Gur RE. Comorbidity of physical and mental disorders in the Neurodevelopmental Genomics Cohort Study. Pediatrics. in press. in press.

Meyerson P, Tryon WW. Validating Internet research: A test of the psychometric equivalence of Internet and in-person samples. Behavior Research Methods, Instruments, & Computers. 2003; 35(4):614–620.

Montse A, Pere V, Carme J, Francesc V, Eduardo T. Visuospatial deficits in parkinsons disease assessed by judgment of line orientation test: Error analyses and practice effects. Journal of Clinical and Experimental Neuropsychology. 2001; 23(5):592–598. [PubMed: 11778636]

Moore TM, Reise SP, Gur RE, Hakonarson H, Gur RC. Psychometric Properties of the Penn Computerized Neurocognitive Battery. Neuropsychology. 2014 doi:10.1037/neu0000093.

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

Rabin LA, Barr WB, Burton LA. 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. 2005; 20(1):33–65. doi:10.1016/j.acn.2004.02.005. [PubMed: 15620813]

Reise SP, Henson JM. Computerization and adaptive administration of the NEO PI R. Assessment. 2000; 7(4):347–364. [PubMed: 11151961]

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

Ritter P, Lorig K, Laurent D, Matthews K. Internet versus mailed questionnaires: a randomized comparison. Journal of Medical Internet Research. 2004; 6(3)

Roalf DR, Ruparel K, Gur RE, Bilker W, Gerraty R, Elliott MA, Gur RC. Neuroimaging predictors of cognitive performance across a standardized neurocognitive battery. Neuropsychology. 2014; 28(2):161–176. doi:10.1037/neu0000011. [PubMed: 24364396]

Roalf DR, Ruparel K, Verma R, Elliott MA, Gur RE, Gur RC. White matter organization and neurocognitive performance variability in schizophrenia. Schizophrenia Research. 2013; 143(1): 172–178. doi:10.1016/j.schres.2012.10.014. [PubMed: 23148898]

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement. 1969; 34(4)(Pt. 2):100.

Satterthwaite TD, Elliott MA, Ruparel K, Loughead J, Prabhakaran K, Calkins ME, Gur RE. Neuroimaging of the Philadelphia Neurodevelopmental Cohort. NeuroImage. 2014; 86:544–553. doi:10.1016/j.neuroimage.2013.07.064. [PubMed: 23921101]

Satterthwaite TD, Wolf DH, Erus G, Ruparel K, Elliott MA, Gennatas ED, Gur RE. Functional maturation of the executive system during adolescence. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience. 2013; 33(41):16249–16261. doi:10.1523/JNEUROSCI.2345-13.2013. [PubMed: 24107956]

Segall, DO. Computerized adaptive testing.. In: Kempf-Leonard, K., editor. Encyclopedia of social measurement. Academic Press; New York, NY: 2005.

Smith GT, McCarthy DM, Anderson KG. On the sins of short-form development. Psychological Assessment. 2000; 12(1):102–111. [PubMed: 10752369]

Spencer RJ, Wendell CR, Giggey PP, Seliger SL, Katzel LI, Waldstein SR. Judgment of Line Orientation: an examination of eight short forms. Journal of Clinical and Experimental Neuropsychology. 2013; 35(2):160–166. doi:10.1080/13803395.2012.760535. [PubMed: 23350928]

Thomas P, Bhatia T, Gauba D, Wood J, Long C, Prasad K, Deshpande SN. Exposure to herpes simplex virus, type 1 and reduced cognitive function. Journal of Psychiatric Research. 2013; 47(11):1680–1685. doi:10.1016/j.jpsychires.2013.07.010. [PubMed: 23920011]

Trahan DE. Judgment of line orientation in patients with unilateral cerebrovascular lesions. Assessment. 1998; 5(3):227–235. [PubMed: 9728030]

Tranel D, Vianna E, Manzel K, Damasio H, Grabowski T. Neuroanatomical correlates of the Benton Facial Recognition Test and Judgment of Line Orientation Test. Journal of Clinical and Experimental Neuropsychology. 2009; 31(2):219–233. doi:10.1080/13803390802317542. [PubMed: 19051129]

Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, WU-Minn HCP Consortium. The Human Connectome Project: a data acquisition perspective. NeuroImage. 2012; 62(4):2222–2231. doi:10.1016/j.neuroimage.2012.02.018. [PubMed: 22366334]

Weiss DJ, Kingsbury G. Application of computerized adaptive testing to educational problems. Journal of Educational Measurement. 1984; 21(4):361–375.

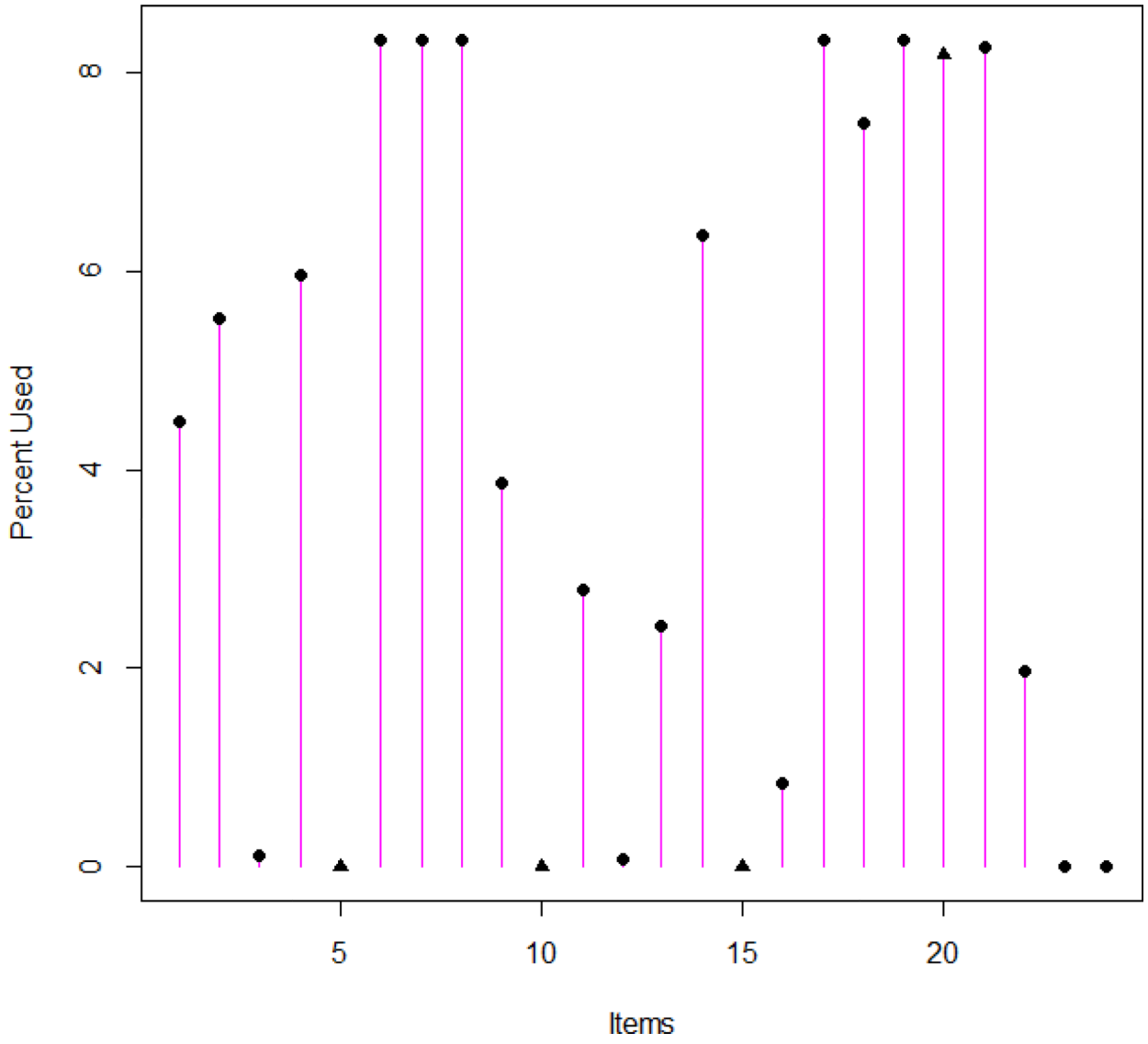

Figure 1.
Histogram of PLOT24 Item Usage for Adaptive Test Simulation of 1000 Examinees and Maximum Item Administration of Twelve per Examinee.

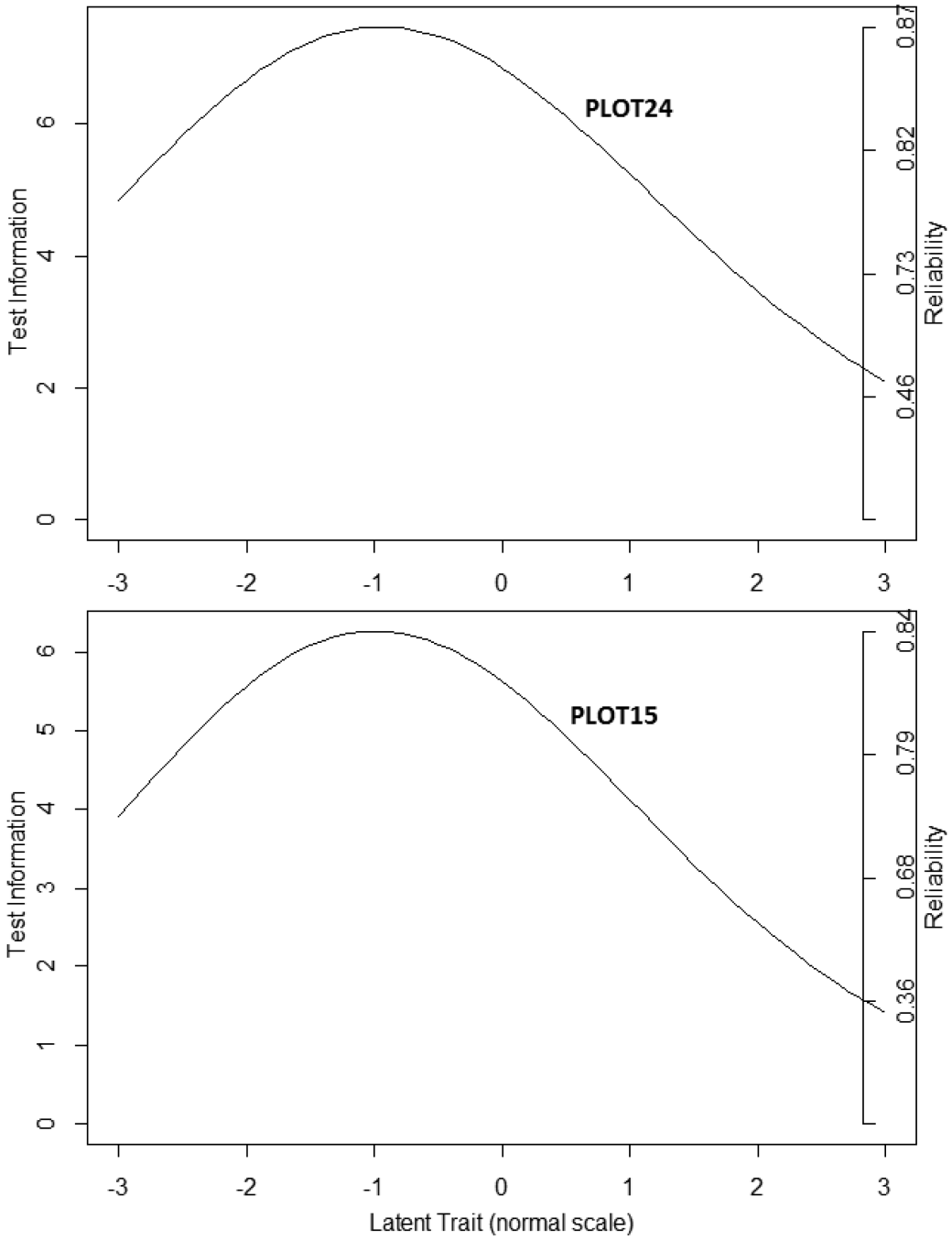

Figure 2.
Test Information Curves for the PLOT24 and its Shortened Version, the PLOT15.

Table 1

Full US sample factor analysis and Graded Response Model (GRM) parameter estimates for the VSPLOT24 with responses coded as polytomous and dichotomous.

Item types			Polytomous					Dichotomous		
Item	Degrees per click	Line length	Factor loading	Disc	Locations (difficulties)			Factor loading	Disc	δ_1
					δ_1	δ_2	δ_3			
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

Psychol Assess. Author manuscript; available in PMC 2016 September 01.

Note. Disc = Discrimination; L = Long; M = Medium; IRT parameters reported in normal units; to convert to conventional IRT parameters, multiply by 1.702; ratios of 1st/2nd eigenvalues for the polytomous and dichotomous models were 5.11 and 3.70, respectively; root mean-square error of approximation (RMSEA) for the polytomous and dichotomous models were 0.058 (± 0.001) and 0.051 (± 0.001), respectively; root mean square of the residuals (RMSR) was 0.04 for both the polytomous and dichotomous models.

Psychol Assess. Author manuscript; available in PMC 2016 September 01.

Table 2

VSPLOT24 *polytomous* item selection based on 1000 simulated CAT sessions using a normal distribution of theta, by sample and item type.

	Item types		Samples							
Item	Degrees Per Click	Line Length	Full U.S. sample	By gender		By age			British sample	Rem.
				Male	Female	8 - 10	11 - 17	18 - 21		
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. L = Long; M = Medium; S = selected; Rem. = removed from final VSPLOT15 battery.

Table 3

VSPLOT24 *dichotomous* item selection based on 1000 simulated CAT sessions using a normal distribution of theta, by sample type.

	Item types		Samples							
Item	Degrees Per Click	Line Length	Full U.S. sample	By gender		By age			British sample	Rem.
				Male	Female	8 - 10	11 - 17	18 - 21		
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. L = Long; M = Medium; S = selected; Rem. = removed from final VSPLOT15 battery.

Psychol Assess. Author manuscript; available in PMC 2016 September 01.

Table 4

Means (and percent) 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		
Score	**Means**		**Diff (M - F)**	**t**	**p-value**	**Corr with age**	**t**	**p-value**
	Male	**Female**						
Full-Length	48.12 (66.8%)	45.46 (63.1%)	2.66 (3.7%)	10.3	< 0.001	0.402	40.0	< 0.001
Shortened	32.07 (71.3%)	30.70 (68.2%)	1.37 (3.1%)	7.4	< 0.001	0.408	40.7	< 0.001

Note. 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.