Social Science Research 41 (2012) 1003–1016

Contents lists available at SciVerse ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Improving ability measurement in surveys by following the principles of IRT: The Wordsum vocabulary test in the General Social Survey

M. Ken Cor [a], Edward Haertel [a], Jon A. Krosnick [b], Neil Malhotra [c,*]

[a] School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA 94305-3096, United States
[b] Departments of Communication and Political Science, Stanford University, Rm. 434 McClatchy Hall, 450 Serra Mall, Stanford, CA 94305, United States
[c] Graduate School of Business, Stanford University, 655 Knight Way, Stanford, CA 94305, United States

ARTICLE INFO

Article history:
Received 10 August 2011
Revised 26 April 2012
Accepted 2 May 2012
Available online 16 May 2012

Keywords:
Item response theory (IRT)
Classical test theory (CTT)
Wordsum
General Social Survey
Vocabulary
Ability
Intelligence
Measurement
Test construction
Scales

ABSTRACT

Survey researchers often administer batteries of questions to measure respondents' abilities, but these batteries are not always designed in keeping with the principles of optimal test construction. This paper illustrates one instance in which following these principles can improve a measurement tool used widely in the social and behavioral sciences: the GSS's vocabulary test called "Wordsum". This ten-item test is composed of very difficult items and very easy items, and item response theory (IRT) suggests that the omission of moderately difficult items is likely to have handicapped Wordsum's effectiveness. Analyses of data from national samples of thousands of American adults show that after adding four moderately difficult items to create a 14-item battery, "Wordsumplus" (1) outperformed the original battery in terms of quality indicators suggested by classical test theory; (2) reduced the standard error of IRT ability estimates in the middle of the latent ability dimension; and (3) exhibited higher concurrent validity. These findings show how to improve Wordsum and suggest that analysts should use a score based on all 14 items instead of using the summary score provided by the GSS, which is based on only the original 10 items. These results also show more generally how surveys measuring abilities (and other constructs) can benefit from careful application of insights from the contemporary educational testing literature.

1. Introduction

Social and behavioral scientists often measure abstract constructs using batteries of survey questions, and the obtained measurements are then combined to yield summary scores for use in statistical analyses. This is often done to measure abilities. For example, a great deal of literature in political science has used survey questions to measure the amount of factual knowledge that respondents possess about politics (e.g., Delli Carpini and Keeter, 1996). Similarly, the National Longitudinal Survey of Youth has administered the Armed Services Vocational Aptitude Battery (ASVAB), which contained ten subtests in 1979 and twelve subtests in 1997 to assess science and vocabulary knowledge, arithmetic reasoning ability, and other individual attributes (Bureau of Labor Statistics, 2005). And since 1974, the General Social Survey (GSS) has measured respondents' vocabulary knowledge with a quiz called "Wordsum" that has been used in numerous research projects in sociology and other disciplines as well.

In this paper, we focus on Wordsum and illustrate how this measurement tool can be improved in a way that is routinely overlooked by survey researchers: optimizing the distribution of item difficulties. For example, nowhere in Delli Carpini and

* Corresponding author.
E-mail addresses: mcor@ualberta.ca (M.K. Cor), haertel@stanford.edu (E. Haertel), krosnick@stanford.edu (J.A. Krosnick), neilm@stanford.edu (N. Malhotra).

0049-089X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ssresearch.2012.05.007

Keeter's (1996) important book on the development of a measure of political knowledge is there a discussion of improving the assessment process by this method. But in the educational testing literature, such optimizing is well recognized as an essential component of effective ability test construction for the purpose of producing scores that reliably and validly rank individuals on the underlying dimension of interest (Hambleton and Jones, 1993). We illustrate how adding well-chosen items to the Wordsum test enhances its measurement of vocabulary knowledge and allows scholars to make better inferences about its relations to other constructs of interest.

2. The Wordsum test and its use in the social, behavioral, and cognitive sciences

Many tests have been constructed to measure vocabulary knowledge, most of them very lengthy. Well-known tests used in educational and psychological research include the vocabulary items of the I.E.R. Intelligence Scale CAVD, the vocabulary subtest of the Wechsler Adult Intelligence Scale-Revised (WAIS-R) (Wechsler, 1981), the Mill-Hill Vocabulary Scale (Raven, 1982), the vocabulary section of the Nelson-Denny Reading Test (Nelson and Denny, 1960), the vocabulary subtest of the Shipley Institute of Living Scale (Shipley, 1946), and others. Some tests (e.g. the items from the I.E.R. Intelligence Scale CAVD) are multiple-choice, whereas others (e.g. WAIS-R) ask respondents to provide open-ended answers. The WAIS-R includes 35 vocabulary items in a 60- to 90-min test; the Mill-Hill scale is composed of 66 questions, entailing 25 min of testing time; the Nelson-Denny test presents 80 vocabulary items in a 45-min test; and the Shipley test includes 40 items in a 20-min assessment.

In contrast to these lengthy measures, the GSS's ten-item, multiple-choice "Wordsum" measure of vocabulary knowledge is much shorter and has been included in twenty surveys of representative national samples of American adults between 1974 and 2010. Wordsum originated in Thorndike's early research on cognitive ability and intelligence testing. In the early 1920s, Thorndike developed a lengthy vocabulary test as part of the I.E.R. Intelligence Scale CAVD to measure, in his words, "verbal intelligence." As in the modern-day Wordsum test, each question asked respondents to identify the word or phrase in a set of five whose meaning was closest to a target word. Thorndike (1942) later extracted two subsets of the original test, each containing twenty items of varying difficulty. For each subset, two target words were selected at each of ten difficulty levels. The ten items in Wordsum (labeled with the letters A though J) were selected from the first of these two subsets.[1]

Wordsum has been administered using a show card that interviewers hand to GSS respondents during interviews in their homes. Each prompt word in capital letters is followed by five response options (as well as a "don't know" option), all numbered and in lower-case. Some response options are single words, while others are phrases.[2] The instructions provided to respondents are:

> "We would like to know something about how people go about guessing words they do not know. On this card are listed some words—you may know some of them, and you may not know quite a few of them.On each line the first word is in capital letters—like BEAST. Then there are five other words. Tell me the number of the word that comes *closest* to the meaning of the word in capital letters. For example, if the word in capital letters is BEAST, you would say '4' since 'animal' comes closer to BEAST than any of the other words.If you wish, I will read the words to you. These words are difficult for almost everyone—just give me your best guess if you are not sure of the answer. CIRCLE ONE CODE NUMBER FOR EACH ITEM BELOW.EXAMPLEBEAST 1. afraid 2. words 3. large 4. animal 5. separate 6. DON'T KNOW"

Wordsum has been used extensively as an independent variable and a dependent variable in much previous research.[3] Between 1975 and 2011, more than 100 studies published in social science journals, books, and edited volumes used Wordsum (for a partial list, see Online Appendix A).[4] The majority of these studies were published in sociology, political science, education, and psychology, though Wordsum has appeared in publications in other disciplines as well.

Pooling together data from the 1974 to 2008 GSSs reveals that Wordsum is solely composed of difficult and easy items. Six of the ten items (A, B, D, E, F, I) were answered correctly by 82%, 90%, 95%, 79%, 79%, and 74% of respondents, respectively, and the remaining four items (C, G, H, J) were answered correctly by only 18%, 31%, 29%, and 24% of respondents, respectively. Hence, the test is missing items answered correctly by between 32% and 73% of respondents.[5]

3. Optimizing test design: principles of item response theory

According to classical test theory (Lord and Novick, 1968) and item response theory (Lord, 1980), the distribution of item difficulties that should be included in a test is a function of the purpose of the test. For example, when scores are used to rank

[1] Prior to its initial use in the 1974 GSS, a slightly different version of Wordsum was used in another national survey: National Opinion Research Center (NORC) Study SRS-889A (1966).

[2] The administrators of the GSS keep the test item wordings confidential to avoid contamination of future surveys, so we cannot present the items' wordings here. Following GSS practice, we refer to the items using the letters A through J, corresponding to their order of administration.

[3] Researchers have often used correlations between Wordsum and other variables to explore the plausibility of causal hypotheses about the origins or consequences of vocabulary knowledge. In this paper, we do not set out to make causal claims and instead simply examine these same sorts of cross-sectional associations.

[4] We assembled this list via Google Scholar using the search terms "General Social Survey AND vocabulary" and "General Social Survey AND Wordsum." We then read each article to determine whether it employed the Wordsum test in a statistical analysis. This approach is likely to have *undercounted* the number of studies that used Wordsum.

[5] In all analyses DK responses are treated as incorrect. See footnote of Table 1 for the definition of missing cases.

order individuals (as is done with Wordsum), a test should be designed to yield a broad range of scores that discriminate validly among examinees as much as possible (Hambleton and Jones, 1993). In other words, tests used to rank-order individuals should have high quality items at most levels of difficulty.

IRT defines how the probability of answering a test question correctly given an examinees' underlying ability can be represented mathematically using latent-trait parametric models. For example, as shown in Eq. (1), the three parameter logistic (3PL) model as specified by Lord (1980) states that the probability of answering a question correctly given the respondent's ability, $p_i(\theta)$, is a function of the item discrimination parameter, a, the item difficulty parameter, b, and the item pseudo-guessing parameter, c, where i indexes the item:

$$p_i(\theta) = c_i + (1 - c_i)/[1 + e_i^{-\alpha(\theta - b)}] \quad (1)$$

In the 3PL model, the discrimination parameter describes the effectiveness of an item in distinguishing between examinees with higher versus lower levels of the ability the test is designed to measure. Higher estimated values of the a-parameter indicate better discrimination. The difficulty parameter identifies where along the underlying ability continuum an item is most discriminating. For example, a large b-parameter indicates that an item is more difficult and is therefore most effective at measuring people of high ability. Finally, the pseudo guessing parameter estimates the probability of answering an item correctly given a very low ability. In other words, it is the probability of a lucky guess.

The standard errors of ability estimates (conditional on true ability) are a function of these three item parameters for all the items included on a test. Conditional standard errors are lowest at the point where the test discriminates most highly. In other words, the lowest conditional standard error for a test is typically found in the region of the ability scale closest to most items' b-parameters. In order to provide the most accurate estimates of ability for most people taking a test, tests should be constructed using items that measure most precisely in the region of the ability scale where most examinees' abilities lie. This makes intuitive sense. A test cannot differentiate accurately among examinees for whom it is too difficult (all of whom will score around chance) nor among those for whom it is too easy (all of whom will get near-perfect scores).

Consequently, experts have long advised that it is especially important to have items that are of moderate difficulty in a normative test setting. For example, according to Bielinksi et al. (2000): "Item selection is driven by the desire to provide precise test score measurement for the majority of the test taking population. This is accomplished by selecting items that are moderately difficult for examinees". Minnema and Thurlow (2000) concur: "Replacing moderately difficult items with easier items mitigates error for a relatively small number of low performing examinees at the expense of a drop in precision for the majority of examinees."

Taken together, IRT implies that Wordsum, as it stands today, is an optimal measure of vocabulary knowledge *only* if most respondents are of extremely low or extremely high ability. However, a quick inspection of the distribution of scores on the Wordsum test for a recent administration of the GSS as shown in Fig. 1 does not seem to support this claim. Instead, it appears that the distribution of vocabulary knowledge is in fact uni-modal, with most respondents scoring between 40 and 70% on the test. Interestingly, the grey areas of the distribution show that only 24% of respondents functioned at ability levels the items are best suited to measure. This distribution of scores suggests that the test's properties might be improved if it were to include items with moderate levels of difficulty.

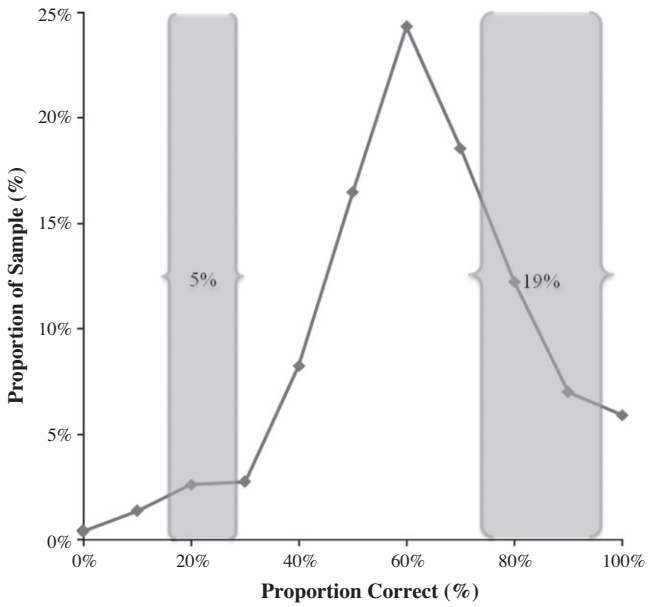

Fig. 1. Observed score distribution for the 2008 administration of the Wordsum test on the GSS.

4. Selection of new words

To explore whether this is true, we set out to identify a set of moderately difficult items to add to Wordsum. A natural place to look for such items is the vocabulary section of the I.E.R. Intelligence Scale CAVD, since Wordsum is itself a subset of this larger test. Because we were unable to locate any CAVD test results from the last few decades, we developed a technique to determine which of the CAVD test words were most likely to be moderately difficult using the frequency of the words' occurrence in popular news media stories.

Our approach was based on the assumption that the more frequently a word is used in news stories, the more likely people are to know its meaning. Such an association between word frequency in news stories and public understanding of the words could result from two phenomena: (1) the news media might avoid using words that people do not understand; and/or (2) people might be more likely to learn the meanings of words to which they are exposed more frequently in news stories. Either way, frequency of appearance in news stories might serve as an indicator of item difficulty.

To test this hypothesis, we began by using Lexis-Nexis to count the number of stories in *The New York Times* that contained each of the ten Wordsum words in the headline or lead paragraph between 1982 and 2000, the years for which data were available. With those data, we estimated the parameters of the following OLS regression equation:

$$Percent\ Correct_i = \beta Ln\ Stories_i + \varepsilon_i. \tag{2}$$

where $Percent\ Correct_i$ is the percent of respondents who correctly answered the Wordsum question about word i, and $Stories_i$ is the number of news stories that included word i in the headline or lead paragraph.

A standardized estimate of the relation between the natural log of the number of stories and the percent correct was $r = .68$ ($R^2 = .46$, $p = .03$), a strong correlation. The unstandardized coefficient is 13.04, meaning that a 1% increase in the number of stories was associated with a .13 percentage point increase in correct responses. This suggests that we could use the frequency of news media mentions of words in the CAVD that are *not* in Wordsum to predict the percent of Americans who would define each word correctly.

To begin the process of selecting candidate items for adding to Wordsum, we randomly selected thirteen test items from the intermediate levels of the CAVD (which are the levels from which the Wordsum items were selected–Levels V3, V4, V5, V6, and V7).[6] We then generated predicted percent correct scores for these words using their frequency in news stories. Seven of the words had predicted percent correct scores between 40% correct and 60% correct. These therefore seemed worthy of further investigation.

We used a second source of information to identify potential words to add as well: the results of tests administered by Thorndike et al. (1927) to high school seniors on various occasions between 1922 and 1925, as described in his book *The Measurement of Intelligence*. Clearly, this respondent pool is very different from a national probability sample of American adults living today. However, the correlation between percent correct for the ten Wordsum words in the 1922–1925 Thorndike sample and the 1974–2000 GSS samples is a remarkable .83, meaning that the difficulty rankings and the differences in difficulties between words were consistent across datasets. Hence, the Thorndike results may offer a useful opportunity to select items for testing with the American public today.

In Thorndike's data, 17 words were correctly defined by between 42% and 62% of high school seniors. One of these words was also identified by our method using news story frequency to estimate item difficulty, making it an especially appealing candidate. Using all the items for which we had predicted percent correct from both our news story frequency analysis and also from Thorndike's testing, the correlation between the sets of predictions was $r = .40$.[7] Thus, there was some correspondence between the two methods for this set of words, though correspondence was far from perfect.

To gauge whether these methods identified test items that would in fact be moderately difficult and therefore useful additions to Wordsum, we administered 23 items from the CAVD (the seven words from the news story analysis and sixteen additional words from the Thorndike administration) to a general population sample of American adults to ascertain the percent of people who answered each one correctly. We also administered the ten Wordsum items to assess comparability of results from this sample to those obtained in the GSS.

The 23 new test items were included in an Internet survey in January, 2007, of a non-probability sample of 1498 American adults who volunteered to complete surveys for Lightspeed Research.[8] The proportions of the Lightspeed respondents answering the ten Wordsum questions correctly were higher than the proportions of GSS respondents doing so, by an average of 7.6% points. However, the ranking of difficulties of the ten Wordsum items was about the same in both surveys. In fact, the correlation between the percent correct across the 10 items was an extraordinary $r = .99$. Hence, results from the Lightspeed survey for the 23 proposed new words seemed likely to be informative about how GSS respondents would answer the items.

[6] Four of these items were eventually included in the 2008 GSS, so we do not describe them here and refer to them anonymously (i.e., K, L, M, N).

[7] This correlation is based on 20 words, because three of the words were not administered by Thorndike et al. (1927).

[8] Lightspeed's panel of potential survey respondents was recruited in three principal ways: (1) people who registered at a website for some non-research purpose and agreed to receive offers from other organizations were later sent emails inviting them to join the Lightspeed panel to complete survey questionnaires, (2) people who registered at a website for some non-research purpose and checked a box at that time indicating their interest in joining the Lightspeed panel were later sent emails inviting them to complete the Lightspeed registration process, and (3) banner advertisements on websites invited people to click and join Lightspeed's panel. Using results from the U.S. Census Bureau's Current Population Survey, Lightspeed Research quota-sampled its panel members in numbers such that the final respondent pool would resemble the US population as a whole in terms of characteristics such as age, gender, and region. Post-stratification weights were constructed so that the sample matched the US population in terms of education, race, age, and gender.

To anticipate the percent correct for these words likely to occur in a representative sample of the general public, we estimated the parameters of an OLS regression predicting the percent of GSS respondents who answered each item correctly using the percent who did so in the Lightspeed survey. The coefficient estimates for the intercept and slope were −6.9 ($p = .16$) and .99 ($p < .001$), respectively. Hence, on average, there was a nearly perfect 1:1 relation between GSS and Lightspeed percents correct and a 6.9% point intercept shift. Correcting for this discrepancy, we used the regression parameters to calculate predicted percent correct values in the GSS for the new test items administered in the Lightspeed survey. According to this method, twelve words manifested predicted percents correct in the moderate range (40–60% correct).

To select the most desirable items, we sought to identify those with the highest discrimination parameters from an IRT analysis. Our first step in this process involved conducting IRT analyses with responses to the ten Wordsum items in the Lightspeed Research dataset. The correlations across the GSS and Lightspeed data were $r = .82$ for the discrimination parameters and $r = .97$ for the difficulty parameters. This, too, inspires some confidence in use of the Lightspeed data for identifying new items. Using all 33 items in the Lightspeed dataset (the 10 Wordsum items and the 23 possible additions), we again estimated a three-parameter IRT model, producing discrimination and difficulty statistics for the proposed additional words. Words with the highest discrimination scores are the most appealing to add to Wordsum–we chose the four highest ones (out of the twelve moderately-difficult items) to administer along with the ten existing items.

5. Data

To evaluate the impact of adding items to Wordsum, we analyzed data from five surveys: the 2008 General Social Survey panel re-interviewing respondents from the 2006 survey (which administered Wordsum plus the four additional items); a wave of data collected on the Face-to-Face Recruited Internet Equipped Survey Platform (FFRISP); two Internet surveys of representative samples of American adults conducted by KnowledgePanel (KP, from the American National Election Studies Internet Panel and the KnowledgePanel); and the Lightspeed Research sample described above. In the analyses below, we compare responses for the 14-item scale to the 10-item scale for the same group of respondents, as Wordsum is simply a subset of Wordsumplus and was always administered before the additional four items.[9] For some analyses, we pooled the data from all surveys together. For each dataset, methods of data collection, sample size, and response rate are described in Table 1.

6. Results

In all of the surveys, many respondents had total test scores in the moderate range (see Fig. 2), and the mean test scores were also in the moderate range (see Table 2). Most importantly, the four new items (words K, L, M, and N) were of moderate difficulty, with percents of respondents answering correctly in the range between the easy and difficult items in the original Wordsum test (see Fig. 3). The middle range of difficulty—between 50% and 75% correct—would have been empty if it were not for the addition of these four items.

To assess whether adding these four items improved the measurement properties of Wordsum, we first tested whether reliability increased when evaluated from the perspectives of Classical Test Theory (CTT) and Item Response Theory (IRT). Then we tested whether concurrent validity improved.

6.1. Classical test theory reliability

Including the four new items increased the reliability of Wordsum estimated by Cronbach's alpha, a commonly used index of CTT reliability that represents a conservative estimate of the overall reliability of a test. Cronbach's alpha for the original ten-item test in the pooled dataset is .678, and adding the four new items increased alpha to .787 (see Table 3). This result is not surprising, because in general, reliability increases when similar items are added to lengthen a test (as described by the Spearman–Brown prophecy formula, see Haertel, 2006). The Spearman–Brown formula's prediction of the reliability of a 14-item test based on the estimated reliability of the 10-item test (.678) for the pooled sample is .747. The difference between the actual reliability of the 14-item test and the Spearman–Brown prediction of reliability ($D = .787 - .747 = .04$) is highly significant ($t = 15.18$; $p < .001$).[10] This suggests that adding the four new items improved Cronbach's alpha more than would be expected simply as the result of adding four either very difficult or very easy items to the original Wordsum test, thereby increasing its length but not changing its distribution of item difficulties. The same pattern appeared in all five of the datasets and was statistically significant in every instance (see Table 3).

[9] In analyzing Wordsumplus, we dropped respondents who did not answer the four additional test items. Given the small number of respondents for which this was the case, comparing Wordsum and Wordsumplus using a common set of respondents yields similar results for all analyses.

[10] The significance of the difference between the actual reliability of the 14-item test and the Spearman-Brown estimate of reliability is calculated using the Delta method (see Papke and Wooldridge, 2005). For this particular case, the procedure involved (1) expressing the reliabilities as functions of the relevant sums of variances and covariances, (2) calculating the covariance matrix of the item variances and covariances, (3) calculating the covariance matrix of the relevant sums of variances and covariances, (4) determining the Jacobian of the relevant sums of variances and covariances, (5) calculating the variance covariance matrix of the reliabilities by pre- and post-multiplying the result of (3) by the matrix produced in (4), and (6) calculating the standard error of the difference by pre- and post- multiplying by the row vector $[1, -1]$ to achieve the appropriate linear combination of variances and covariances.

Table 1
Details of samples used for analyses.

	GSS	FFRISP	ANES	KP	Lightspeed
Sample size	1536	981	1397	1210	1498
Wordsum sample[a]	739	979	1397	1207	1498
Wordsumplus sample[b]	727	975	1395	1199	1498
Survey field dates	April–September, '08	December '08–January '09	August–September '08	August–September '08	January '07
Response rate	71%	42%	25%	NA	NA
Completion rate	NA	NA	NA	71%	17%
Sample drawn by	NORC	SRC	KP	KP	LR
Data collection	NORC	Abt SRBI Inc.	KP	KP	LR
Sampling method	Area probability	Area probability	RDD	RDD	Non-probability
Recruitment	Face to face	Face to face	Telephone	Telephone	Internet Opt-in
Interview mode	Face to face	Internet	Internet	Internet	Internet
Over sampled minorities	No	No	Yes	Yes	No
Drawn from larger panel	No	No	No	Yes	Yes
Unequal probability of invitation	No	No	No	Stratification with demos	Stratification with demos
Start-up incentives	None	Laptop or $500 + Internet	MSNTV or Cash	MSNTV or cash	None
Incentives for each survey	$0-$100	$5 or $4	$10, $25 or $50	Cash/sweepstake	Points

Note. GSS – General Social Survey, FFRISP – Face-to-Face Recruited Internet Survey Platform, ANES – American National Election Studies Internet Panel, KP – KnowledgePanel, Lightspeed – Lightspeed Research Survey; NORC – National Opinion Research Center at University of Chicago, SRC – Survey Research Center at University of Michigan, KP – KnowledgePanel, LR – Lightspeed Research; RDD – Random Digit Dial.

[a] Any person not administered the Wordsum items or who chose not to respond to more than five consecutive questions is not a part of the Wordsum sample.

[b] Any person who chose not to respond to the four additional wordsum items is not a part of the Wordsumplus sample.

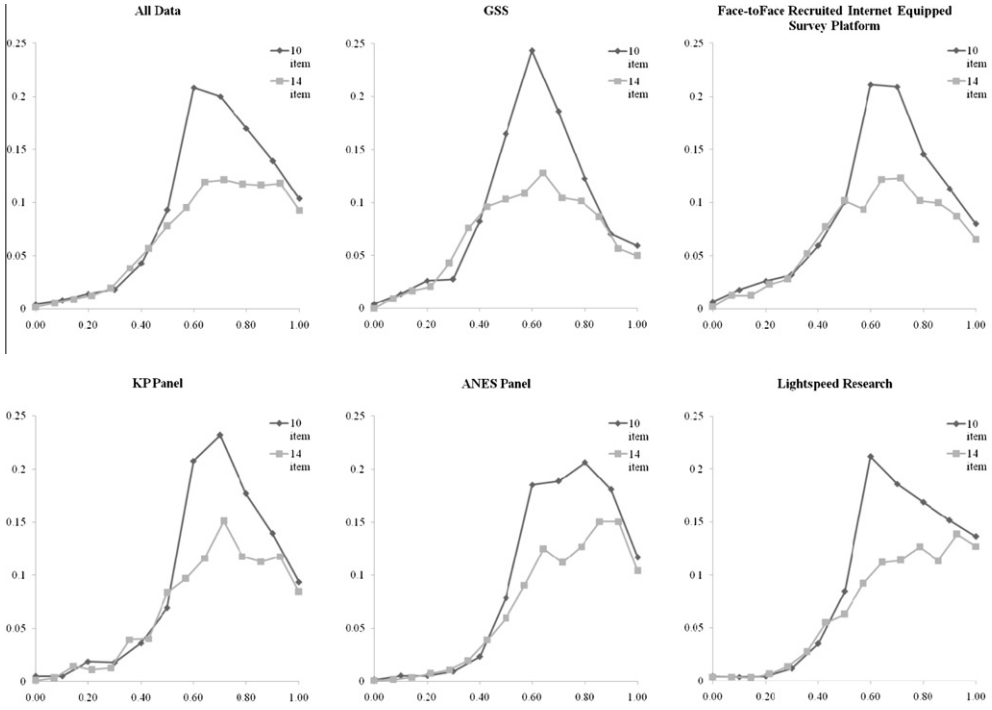

Fig. 2. Observed score distributions as a function of test length across samples.

6.2. Item response theory and conditional standard error

To determine how the reliability of the test varied across the range of underlying ability levels, we applied principles of IRT. Using the IRT software BILOG-MG (Zimowski et al., 1996), we estimated the parameters of a 3-PL model for the pooled sample.[11] The *a*-parameter, which is the discrimination parameter, provides information about how well each item discriminated

[11] Parameters were estimated via marginal maximum likelihood. We assumed a fixed normal prior for the distribution of the ability parameters (the BILOG-MG default option).

Table 2
Mean, proportion correct, and standard deviation as a function of test length and data source.

		10 Items		14 Items	
Sample	*n*	Mean	St. dev.	Mean	St. dev.
GSS	727	.627	.196	.615	.216
FFRISP	975	.647	.193	.633	.212
KP	1199	.706	.188	.701	.203
ANES	1395	.741	.175	.743	.187
Lightspeed research	1498	.727	.187	.729	.204
All data	5794	.705	.190	.700	.206

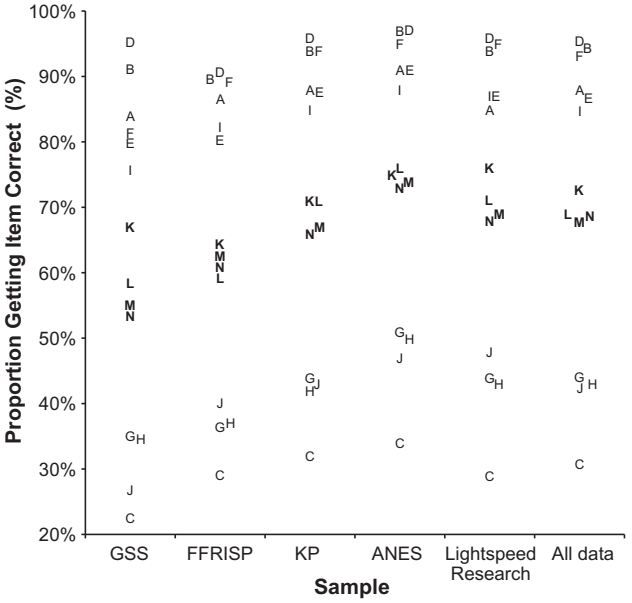

Fig. 3. Proportion correct for each item.

between respondents who had the knowledge required to answer an item correctly and those who did not. Larger *a*-parameters indicate better discrimination. The *b*-parameter is the difficulty parameter; smaller difficulty parameters indicate easier items.

The four new items discriminated well; their *a*-parameters are all larger than the average *a*-parameter for the 14 items (see the second column of Table 4). Furthermore, the *b*-parameters indicate that the new items were of moderate difficulty, with estimates ranging between −.194 and −.017 (clustering around zero indicates that these are moderately difficult items). Taken together, these findings suggest the new items did a better than average job at discriminating in an area of the underlying ability distribution where most examinees are located, but where the original items did not function well. Thus, the four new Wordsum items filled in the difficulty gap in the original 10-item Wordsum test.

The same conclusion is reinforced by the Item Characteristic Curves (ICCs) generated using the merged dataset for the 14-item Wordsum test (see Fig. 4). The further the inflection point of an ICC is to the left on the latent ability scale, the easier the item. The steeper the slope in the middle of the curve, the more discriminating the item. The ICCs for the four new items are centered in the middle of the underlying dimension (shown by solid lines in Fig. 4), surrounded by the ICCs for the original ten items (shown as dotted lines in Fig. 4). If the four new items were removed, a large gap between easy and hard items would be apparent.

The 14-item battery produced lower conditional standard errors for people in the middle of the underlying ability scale—precisely where the 10-item scale is deficient due to the paucity of moderately difficult items. In Fig. 5, the solid line displays the conditional standard error as a function of a person's estimated underlying ability using the ten-item test; the surge in the standard error near the middle of the range is undesirable. The dotted line in Fig. 5, which displays the standard error for the 14-item test, is notably lower in the middle region of the underlying dimension. The same patterns are apparent in each of the datasets separately (see Fig. 6).[12] Taken together, the results show that adding the moderately difficult items pro-

[12] In order to facilitate meaningful comparisons across the datasets, the *c*-parameters generated from the analysis of the pooled dataset were used to fix the *c*-parameters for the analyses of the separate samples. Furthermore, because item parameter estimates are scale indeterminate, all parameters were adjusted to be on a common scale. The mean sigma method (see Kolen and Brennan, 2004) was used to place all parameter estimates on the scale of the FFRISP sample—the selection of scale being arbitrary given that this was done only to facilitate meaningful comparisons across samples.

Table 3
Classical test theory reliability as function of test length and data source.

Sample	*n*	10 Item Cronbach's α (r10)	14 Item S–B prophecy (rSB)	14 Item Cronbach's α (r14)	Difference (r14-rSB)
All data	**5794**	**.678**	**.747**	**.787**	**.040****
GSS	727	.654	.725	.777	.052**
FFRISP	975	.706	.770	.789	.019*
KP	1199	.658	.730	.770	.040**
ANES	1395	.619	.694	.747	.053**
Lightspeed Research	1498	.684	.752	.799	.047**

* $p < .05$ (two-tailed).
** $p < .01$.

duces a test that more precisely measures the abilities of respondents with vocabulary knowledge in the area of ability scale where most respondents were located.

6.3. Concurrent validity

To assess concurrent validity,[13] we first identified all variables that were measured in the GSS, FFRISP, KP, and ANES that correlated at least .15 with either the 10-item Wordsum test score or the 14-item Wordsumplus test score.[14] Twenty variables met this criterion (see Appendix B for the wordings of these questions). We computed the correlations of these variables with Wordsum and Wordsumplus test scores and assessed the statistical significance of the difference between each such pair of correlations, taking into account the partial dependence of the correlations, since they are computed using data from the same respondents.[15]

As expected, the 14-item scale manifested greater concurrent validity than the 10-item scale. This finding is consistent with the conclusion that the 14-item scale exhibited greater reliability than the 10-item scale. Out of 20 tests, Wordsumplus produced more positive correlations in every instance, and in 18 of the 20 cases, the magnitude of the increase was statistically significant ($p < .01$; see Table 5). A sign test also confirmed that this pattern would be extremely unlikely to have occurred by chance alone ($p < .001$). Of the two non-significant differences, both were in the expected, positive direction. Therefore, including the four new items increased the likelihood of detecting more theoretically-sensible correlates of vocabulary skill.

7. Discussion

In sum, CTT reliability indexes, IRT conditional standard errors, and concurrent validity coefficients for the Wordsum and Wordsumplus tests demonstrated that complying with the principles of sound test construction—including items at all ranges of difficulty—produced a better functioning vocabulary knowledge measure. This evidence suggests that the widely-used current version of the Wordsum test is less effective than an expanded test with four additional items of moderate difficulty. The expanded test was especially more accurate in assessing people whose ability levels are most common—in the middle range of the underlying dimension. Replication of these findings across many independent datasets confirms their robustness.

These findings resonate with the widely-accepted principle in educational testing that when a test is used to rank order individuals, the items should be designed to yield a broad range of scores that discriminate among examinees as much as possible (Hambleton and Jones, 1993). The current version of Wordsum can be improved by meeting this widely accepted standard. It therefore seems that researchers interested in a more effective measure of vocabulary knowledge should use Wordsumplus. Based on the findings reported here, GSS users should compute new total test scores using the four additional items when conducting analyses.

[13] Wordsum and Wordsumplus are presumed to measure the same underlying construct: vocabulary knowledge. Therefore, if we were to correct validity correlations for attenuation due to unreliability, we would expect to observe the same magnitudes of association of a criterion with Wordsum and Wordsumplus. Consequently, we did not implement such a correction, so that we could assess whether the lower reliability of the Wordsum measure led to observing weaker associations of it with criteria, compared to associations between the criteria and Wordsumplus.

[14] The Lightspeed Research sample was not used in these analyses because respondents were not asked the same additional questions as in the other four surveys. The results are robust to using various cutoffs for the minimum correlation between items.

[15] No off-the-shelf statistic existed for testing the significance of a difference between correlations with partial dependence. We therefore derived this statistic analytically. The resulting equation is:

$$H_{Zi} = \frac{(r_{10,Zi} - r_{14,Zi})}{\sqrt{\frac{1}{n}\left[(1-r_{10,Zi}^2)^2 + (1-r_{14,Zi}^2)^2 - r_{10,Zi}r_{14,Zi}(r_{10,Zi}^2 + r_{14,Zi}^2 + r_{10,14}^2 - 1) - 2\left(\frac{s_{10}}{s_{14}} + \frac{s_4}{s_{14}}r_{10,4}\right)(1 - r_{10,Zi}^2 - r_{14,Zi}^2)\right]}},$$

where i indexes the criterion variable, Z; $r_{10,Zi}$ is the correlation between the criterion variable and the 10-item test score, $r_{14,Zi}$ is the correlation between the criterion variable and the 14-item test score, $r_{10,14}$ is the correlation between the 10- and 14-item tests, s_{10} is the standard deviation of the scores on the 10-item test, s_{14} is the standard deviation of the scores on the 14-item test, and s_4 is the standard deviation of the scores on the 4-item addendum to the 10-item test. A more generalized form of the equation along with a derivation is provided in Appendix B.

Table 4
3-PL item parameters for the Wordsum test based on the pooled data set.

Word	Proportion correct (p)	Item discrimination (a)	Item difficulty (b)	Guessing parameter (c)
A	.891	.689	−1.784	.116
D	.954	1.410	−1.737	.100
I	.946	1.746	−1.519	.070
B	.841	.626	−1.480	.093
F	.925	1.672	−1.255	.182
E	.869	1.259	−1.084	.080
K	**.719**	**1.481**	**−.194**	**.199**
N	**.688**	**1.433**	**−.111**	**.179**
M	**.681**	**1.560**	**−.041**	**.202**
L	**.672**	**1.425**	**−.017**	**.201**
J	.413	1.351	.639	.071
H	.424	1.767	.729	.145
G	.441	1.538	.759	.174
C	.315	.970	1.158	.082
Average	.700	1.352	−.424	.135

Note. n = 5794; the items are ordered according to the item difficulty parameter; the new words are in bold.

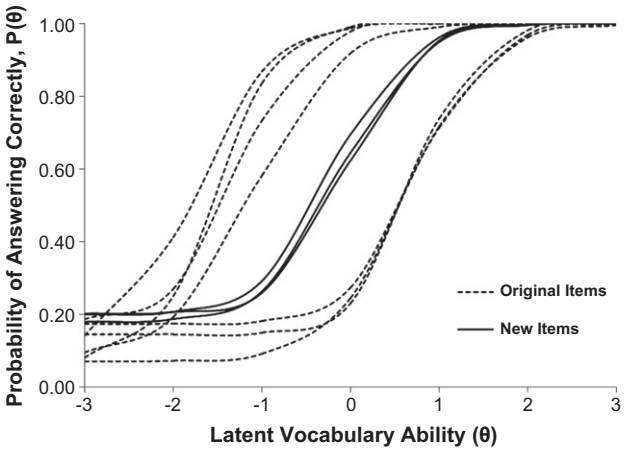

Fig. 4. Item characteristic curves generated from the pooled sample for the 14 Wordsum items.

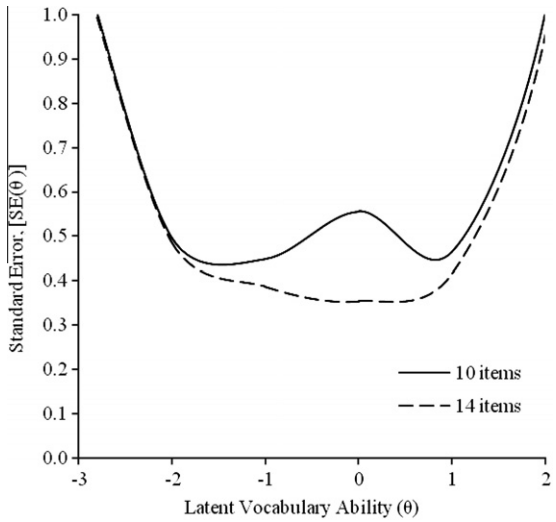

Fig. 5. Conditional standard error as a function of the latent trait of vocabulary knowledge for the pooled dataset.

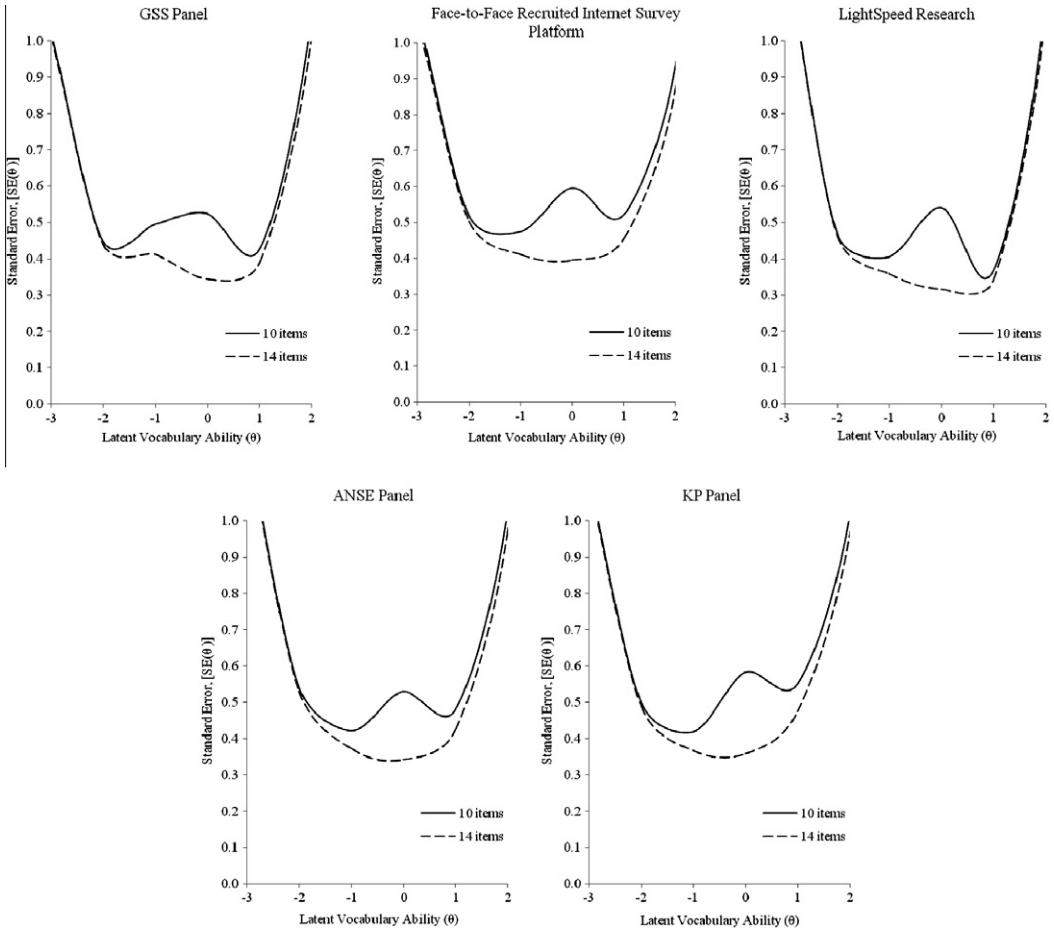

Fig. 6. Conditional standard error as a function of the latent trait of vocabulary knowledge for separate samples.

Table 5
Strength of the relation between total score on the Wordsum tests and selected validity criteria as a function of the Wordsum test length.

Criterion (Z)	n	$r_{10,Z}$	$r_{14,Z}$	diff	β_{10}	β_{14}
Age	4269	.140**	.155**	.015**	.139***	.143***
Education	4284	.210**	.235**	.025***	.350***	.362***
Income	4124	.153**	.171**	.018***	.195***	.202***
Support for preferential hiring of women	3512	.207**	.253**	.046***	.319***	.323***
Anti-religious people should be allowed to teach	4012	.173**	.180**	.007	.422***	.406***
Books against religion should be allowed in libraries	4022	.255**	.287**	.032***	.528***	.550***
Books stating Blacks are less able should be allowed in libraries	4028	.193**	.221**	.028***	.464***	.493***
Self reported SES	4265	.228**	.253**	.025***	.253***	.260***
Belief Blacks don't have less in-born ability than Whites	4004	.188**	.192**	.004	.202***	.192***
Belief that life is exciting	4045	.159**	.175**	.016**	.227***	.232***
Same-sex relations between two adults is okay	3997	.180**	.197**	.017**	.418***	.425***
Spanking children is wrong	4011	.149**	.163**	.014**	.237***	.241***
Lack of confidence in heads of organized labor	3994	.118**	.153**	.035***	.182***	.218***
Lack of confidence in people running TV	4001	.144**	.178**	.034***	.230***	.263***
Confidence in scientific community	3981	.181**	.210**	.029***	.282***	.302***
In general, you can trust people	4024	.218**	.252**	.034***	.287***	.306***
Learning to obey should be a low parenting priority	3281	.255**	.286**	.031***	.342***	.352***
Belief that parents should support independent thinking	3901	.235**	.265**	.030***	.344***	.361***
Support for euthanasia	3988	.165**	.179**	.014**	.424***	.426***
Income separation is necessary for American prosperity	3519	.196**	.214**	.007**	.276***	.281***
Average		.187	.211	.023	.306	.317

Note. $r_{10,Z}$ = correlation between criterion and the 10 item version of the test; $r_{14,Z}$ = correlation between criterion and the 14 item version of the test; diff = difference between $r_{14,Z}$ and $r_{10,Z}$; In order to make comparative interpretations consistent, all criterion variables are coded to reveal a positive relationship with the Wordsum scores.

** $p < .01$ (two-tailed).

*** $p < .001$.

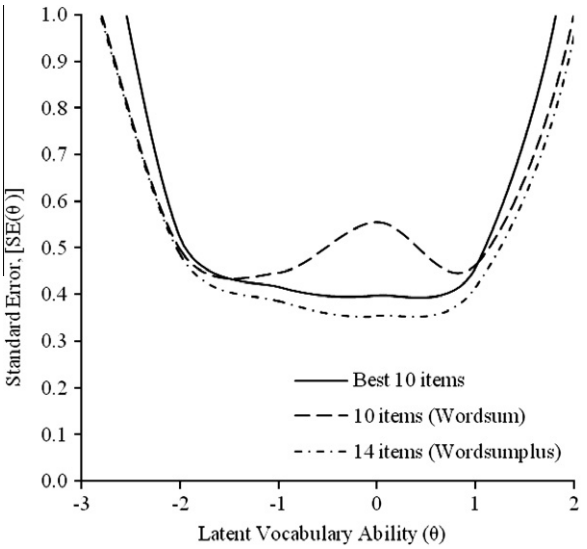

Fig. 7. Comparing the conditional standard error of Wordsum tests consisting of the 10 best items, the original wordsum items, and Worsumplus.

The question might arise as to whether the improvements demonstrated in reliability, conditional standard errors, and concurrent validity were in fact due to the change in the distribution of item difficulties, or were simply due to the inclusion of more discriminating items. In fact, item difficulty and discrimination are interrelated. From a CTT perspective, discrimination statistics typically express one of several kinds of correlation between item response and total test score. These correlations will be low if the item is too easy or too difficult for the group tested. Thus, appropriate difficulty is a prerequisite to high discrimination. From an IRT perspective, an item is regarded as discriminating more or less effectively at different ability levels. Any given item is most discriminating at or slightly above[16] the point on the ability scale where the conditional probability of getting that particular item correct is around .50. Thus, from an IRT perspective as well, sound test design requires inclusion of items of moderate difficulty for the group tested that also have high discrimination parameters.

More broadly, this study provides an illustration of why the measurement properties of longstanding and widely used tests of abilities should be continually revisited. Too many researchers assume that established scales that have been used widely will continue to provide the most reliable and valid measurements of a construct available in subsequent years. In fact, the psychometric properties of all scales should be re-evaluated intermittently so that changes can be made based on sound test construction practices to maintain the fidelity of resulting scores. Researchers would be wise to implement the assessment techniques used in this investigation as a part of this evaluative process.

Of course, there are also tradeoffs associated with adding to or changing the items within scales, especially for long-running time series studies such as the GSS. For example, adding items would allow researchers to continue to make over-time comparisons in the level of vocabulary ability, but come at the cost of increased expense, administration time, and respondent fatigue.[17] Alternatively, these costs could be avoided by choosing the 10 best items out of the 14 in order to replace some of the hard and easy items with moderately-difficult ones. We ran an optimization model based on the goal of minimizing the conditional standard error across the ability scale (see Fig. 7).[18] This optimized scale excluded items A, C, I, and M (one of the new, medium-difficulty items).[19] Compared to the existing 10-item scale (Cronbach's $\alpha = .687$), the optimized 10-item test is much more reliable (Cronbach's $\alpha = .749$).[20] In other words, a substantial improvement in reliability is possible without changing the length of the test. However, replacing hard and easy items with moderately-difficulty ones would disrupt the ability to continue to make over-time comparisons in the level of vocabulary ability. This is not to say that researchers should not improve existing scales, but simply to point out that such improvements require researchers to choose among: (1) spending re-

[16] Discrimination is reflected in slope of the ICC, which reaches its maximum at the ability level where the difficulty is $.50 + c/2$, where c is the pseudo-guessing parameter (i.e., the lower asymptote of the ICC).

[17] Because all of our respondents answered fourteen questions, we were not able to assess whether respondent fatigue was increased by adding four additional items. This could be investigated via a between-subjects experiment, randomly assigning different groups of respondents to complete tests of differing lengths.

[18] Automated Test Assembly (ATA) (see van der Linden, 2005) was used to select the 10 items that minimize the conditional standard error across the score scale. ATA is a method whereby linear optimization is used to select the best combination of items to meet a specific goal, usually defined as some required psychometric characteristic of the test.

[19] These four items did not exhibit the lowest levels of discrimination. Therefore, simply removing the four least-discriminating items does not minimize the conditional standard error across the entire scale. It is important to include items at various levels of difficulty. Note that the best 10-item index includes four easy items, three medium-difficulty items, and three hard items.

[20] The optimized reliability is also about as strong as the Spearman-Brown prediction of the reliability of the 14-item test (see Table 3).

sources to increase scale reliability and validity by expanding to 14 items; (2) increasing scale reliability and validity while preserving resources by using the revised 10-item scale at the cost of disrupting a time series; and (3) keeping the original 10-item scale at the cost of forgoing potential increases in reliability and validity.

Finally, although the 14-item scale was more reliable and valid than the 10-item scale, we may not have constructed the best 14-item scale. The four added items were based on the Thorndike items so they would be as similar as possible to the existing Wordsum items. However, the Thorndike items were not randomly selected from the universe of all possible items, and Thorndike's selection criteria are unclear. Ideally, we would have tested a broader set of items to determine which ones were moderately difficult, highly discriminating, and maximally valid. If anything, our findings understate the potential gains that can be made by applying IRT principles to scale construction in this arena.

The approach employed here to investigate the effects of adding items to the Wordsum test can be generalized to other tests. That is, all of the techniques used to assess how the psychometric properties of the Wordsum test changed with the inclusion of specific additional items can be applied to investigate how adding items to other established batteries improves their measurement in terms of Classical Test Theory outcomes as well as Item Response Theory metrics. We look forward to seeing such work done in the future.

Acknowledgments

Jon Krosnick is University Fellow at Resources for the Future. The authors thank Tom Smith of NORC for extensive help and thank the GSS Board of Overseers for commissioning this research.

Appendix A

Scaled 3-PL item parameters for the Wordsum test across five individual data samples.

Word	*a*-Parameters					*b*-Parameters					*c*-Parameters[a]
	GSS	FFRISP	KP	ANES	LS	GSS	FFRISP	KP	ANES	LR	
A	.82	.73	.72	.65	.66	−1.60	−1.76	−1.67	−1.82	−2.04	.12
D	1.96	1.27	1.29	1.34	1.82	−1.92	−1.65	−1.84	−1.62	−1.56	.10
B	1.37	1.68	2.14	1.67	1.83	−1.75	−1.46	−1.37	−1.50	−1.63	.07
I	.71	.68	.70	.53	.64	−1.23	−1.55	−1.43	−1.83	−1.36	.09
F	1.63	1.31	1.72	1.83	1.91	−.95	−1.43	−1.37	−1.10	−1.30	.18
E	1.00	1.13	1.23	1.35	1.63	−1.24	−1.08	−1.18	−.98	−.93	.08
K	1.42	1.46	1.44	1.51	1.64	−.42	−.17	−.22	−.10	−.19	.20
N	1.48	1.23	1.60	1.50	1.71	−.02	−.04	.03	.01	−.31	.18
L	1.72	1.42	1.51	1.59	1.85	−.08	.02	−.12	−.05	.12	.20
M	1.40	.97	1.39	1.37	1.91	.00	−.10	−.08	−.09	.09	.20
J	1.83	.97	1.38	1.28	1.70	.65	.53	.56	.62	.68	.07
H	1.57	1.31	1.35	1.75	2.71	.65	.79	.75	.63	.75	.15
G	1.93	1.50	1.19	1.57	1.83	.76	.87	.83	.73	.69	.17
C	.91	1.02	.74	.96	1.17	1.21	1.06	1.17	1.16	1.04	.08

[a] *Note*: In order to make meaningful comparisons, *c*-parameters were set to be the same across all samples and the a- and *b*-parameters were scaled to the FFRISP scale. GSS – General Social Survey, FFRISP – Face to Face Recruited Internet equipped Survey Protocol, KP – KnowledgePanel ANES – American National Election Survey, LR – Lightspeed Research. Items are ordered from smallest (easiest) to largest (hardest) difficulty parameter for the merged data set.

Appendix B. A test statistic for the difference between two correlations with partial dependence

Let $x_1, x_2,$ and x_3 be three variables with a multivariate normal distribution. Consider the problem of testing the statistical significance of the difference between the correlation of x_1 with x_2, denoted $\rho_{1,2}$, and the correlation of x_1 with $x_2 + x_3$, denoted $\rho_{1,2+3}$. A suitable test statistic will be

$$(\hat{\rho}_{1,2} - \hat{\rho}_{1,2+3})/\sqrt{\text{var}(\hat{\rho}_{1,2} - \hat{\rho}_{1,2+3})} = (r_{1,2} - r_{1,2+3})/\sqrt{\text{var}(r_{1,2}) + \text{var}(r_{1,2+3}) - 2\text{cov}(r_{1,2}, r_{1,2+3})}$$

The large-sample formula for the variance of a sample correlation is well known, multiplied here by N to simplify further notation:

$$N\ \text{var}(r_{12}) = (1 - \rho_{12}^2)^2$$
$$N\ \text{var}(r_{1,2+3}) = (1 - \rho_{1,2+3}^2)^2$$

The large-sample formula for $N\text{cov}(r_{1,2}, r_{1,2+3})$ is derived in two steps. First, $\rho_{1,2}$ and $\rho_{1,2+3}$ are expressed in terms of the variances and covariances of x_1, x_2 and x_3, denoted σ_1^2, σ_2^2, σ_3^2, σ_{12}, σ_{13}, and σ_{23}. Next, the *Delta method* (e.g. Papke and Wooldridge, 2005) is used to approximate $N\text{cov}(r_{1,2}, r_{1,2+3})$ as $\boldsymbol{\gamma}'_{1,2}\boldsymbol{\Sigma}\boldsymbol{\gamma}_{1,2+3}$ where $\boldsymbol{\gamma}_{1,2}$ is the column vector of partial derivatives of $\rho_{1,2}$ with respect to σ_1^2, σ_2^2, σ_3^2, σ_{12}, σ_{13}, and σ_{23}; $\boldsymbol{\gamma}_{1,2+3}$ is the column vector of partial derivatives of $\rho_{1,2+3}$ with respect to σ_1^2, σ_2^2, σ_3^2, σ_{12}, σ_{13}, and σ_{23}; and $\boldsymbol{\Sigma}$ is N times the covariance matrix of $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, $\hat{\sigma}_3^2$, $\hat{\sigma}_{12}$, $\hat{\sigma}_{13}$, and $\hat{\sigma}_{23}$.

Step 1:

$$\rho_{1,2} = \sigma_{1,2}/\sqrt{\sigma_1^2\sigma_2^2} \text{ and } \rho_{1,2+3} = (\sigma_{1,2}+\sigma_{1,3})/\sqrt{\sigma_1^2(\sigma_2^2+2\sigma_{2,3}+\sigma_3^2)}$$

Step 2:

The elements of $\boldsymbol{\Sigma}$ are each given by the general formula, $N\text{cov}(\hat{\sigma}_{i,j}, \hat{\sigma}_{k,l}) = \sigma_{i,k}\sigma_{j,l} + \sigma_{i,l}\sigma_{j,k}$. Some examples of the elements of $\boldsymbol{\Sigma}$ include $N\text{var}(\hat{\sigma}_1^2) = N\text{cov}(\hat{\sigma}_1^2, \hat{\sigma}_1^2)$, which is expressed as $N\text{cov}(\hat{\sigma}_{1,1}, \hat{\sigma}_{1,1}) = 2(\sigma_1^2)^2$; $N\text{var}(\hat{\sigma}_{1,2}) = (\sigma_{1,2})^2 + \sigma_1^2\sigma_2^2$; $N\text{cov}(\hat{\sigma}_1^2, \hat{\sigma}_2^2) = 2(\sigma_{1,2})^2$; $N\text{cov}(\hat{\sigma}_1^2, \hat{\sigma}_{2,3}) = 2\sigma_{1,2}\sigma_{1,3}$; and $N\text{cov}(\hat{\sigma}_{1,2}, \hat{\sigma}_{1,3}) = \sigma_1^2\sigma_{2,3} + \sigma_{1,2}\sigma_{1,3}$.

Denoting $\sigma_2^2 + 2\sigma_{2,3} + \sigma_3^2$ by σ_{2+3}^2, the elements of $\boldsymbol{\gamma}_{1,2}$ and $\boldsymbol{\gamma}_{1,2+3}$ are as follows:

$$\boldsymbol{\gamma}_{1,2} = \begin{bmatrix} \frac{\partial r_{1,2}}{\partial \sigma_1^2} \\ \frac{\partial r_{1,2}}{\partial \sigma_2^2} \\ \frac{\partial r_{1,2}}{\partial \sigma_3^2} \\ \frac{\partial r_{1,2}}{\partial \sigma_{1,2}} \\ \frac{\partial r_{1,2}}{\partial \sigma_{1,3}} \\ \frac{\partial r_{1,2}}{\partial \sigma_{2,3}} \end{bmatrix} = \begin{bmatrix} -\frac{\sigma_{1,2}}{2\sigma_1^2\sqrt{\sigma_1^2\sigma_2^2}} \\ -\frac{\sigma_{1,2}}{2\sigma_2^2\sqrt{\sigma_1^2\sigma_2^2}} \\ 0 \\ \frac{1}{\sqrt{\sigma_1^2\sigma_2^2}} \\ 0 \\ 0 \end{bmatrix} \text{ and } \boldsymbol{\gamma}_{1,2+3} = \begin{bmatrix} \frac{\partial r_{1,2+3}}{\partial \sigma_1^2} \\ \frac{\partial r_{1,2+3}}{\partial \sigma_2^2} \\ \frac{\partial r_{1,2+3}}{\partial \sigma_3^2} \\ \frac{\partial r_{1,2+3}}{\partial \sigma_{1,2}} \\ \frac{\partial r_{1,2+3}}{\partial \sigma_{1,3}} \\ \frac{\partial r_{1,2+3}}{\partial \sigma_{2,3}} \end{bmatrix} = \begin{bmatrix} -\frac{\sigma_{1,2}+\sigma_{1,3}}{2\sigma_1^2\sqrt{\sigma_1^2\sigma_{2+3}^2}} \\ -\frac{\sigma_{1,2}+\sigma_{1,3}}{2\sigma_{2+3}^2\sqrt{\sigma_1^2\sigma_{2+3}^2}} \\ -\frac{\sigma_{1,2}+\sigma_{1,3}}{2\sigma_{2+3}^2\sqrt{\sigma_1^2\sigma_{2+3}^2}} \\ \frac{1}{\sqrt{\sigma_1^2\sigma_{2+3}^2}} \\ \frac{1}{\sqrt{\sigma_1^2\sigma_{2+3}^2}} \\ -\frac{\sigma_{1,2}+\sigma_{1,3}}{\sigma_{2+3}^2\sqrt{\sigma_1^2\sigma_{2+3}^2}} \end{bmatrix}$$

Substituting, solving, and simplifying,

$$N\ \text{cov}(r_{1,2}, r_{1,2+3}) = \gamma'_{1,2}\sum\gamma_{1,2+3} = \frac{1}{2}\rho_{12}\rho_{1,2+3}\left(\rho_{12}^2 + \rho_{1,2+3}^2 + \rho_{2,2+3}^2 - 1\right) + \left(\frac{\sigma_2}{\sigma_{2+3}} + \frac{\sigma_3}{\sigma_{2+3}}\rho_{23}\right)(1 - \rho_{12}^2 - \rho_{1,2+3}^2)$$

Again substituting and solving, the final test statistic becomes

$$H = \frac{\sqrt{N}(r_{1,2} - r_{1,2+3})}{\sqrt{(1-\rho_{12}^2)^2 + (1-\rho_{1,2+3}^2)^2 - \rho_{12}\rho_{1,2+3}(\rho_{12}^2 + \rho_{1,2+3}^2 + \rho_{2,2+3}^2 - 1) - 2(\frac{\sigma_2}{\sigma_{2+3}} + \frac{\sigma_3}{\sigma_{2+3}}\rho_{23})(1 - \rho_{12}^2 - \rho_{1,2+3}^2)}}$$

Sample values are substituted for parameters. H may be interpreted as a z statistic, referenced to the standard normal distribution. One-tailed or two-tailed tests may be performed.

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ssresearch.2012.05.007.

References

Bielinksi, J., Thurlow, M., Minnema, J., Scott, J., 2000. How Out-of-Level Testing Affects the Psychometric Quality of Test Scores (Out-of-Level Testing Project Report 2). University of Minnesota, National Center on Educational Outcomes, Minneapolis, MN. <http://education.umn.edu/NCEO/OnlinePubs/OOLT2.html> (accessed 09.08.11).

Bureau of Labor Statistics, 2005. National Longitudinal Surveys Handbook. U.S. Government Printing Office, Washington, DC.

Delli Carpini, M.D., Keeter, S., 1996. What Americans Know About Politics and Why It Matters. Yale University Press, New Haven, CT.

Haertel, E.H., 2006. Reliability. In: Brennan, R.L. (Ed.), Educational Measurement, fourth ed. American Council on Education/Praeger, Westport, CT, pp. 65–110.

Hambleton, R.K., Jones, R.W., 1993. Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice 12 (2), 38–47.

Kolen, M.J., Brennan, R.L., 2004. Test Equating, Scaling, and Linking: Methods and Practices, second ed. Springer-Verlag, New York.

Lord, F.M., 1980. Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum, Hillsdale, NJ.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Minnema, J., Thurlow, M., Bielinksi, J., Scott., J., 2000. Past and Present Understandings of Out-of-Level Testing: A Research Synthesis (Out-of-Level Testing Project Report 1). University of Minnesota, National Center on Educational Outcomes, Minneapolis, MN. <http://education.umn.edu/NCEO/OnlinePubs/OOLT1.html>.

National Opinion Research Center, 1966. Study SRS-889A. NORC, Chicago.

Nelson, M.J., Denny, E.C., 1960. The Nelson-Denny Reading Test, Revised by James I. Brown. Houghton Mifflin, Boston.

Papke, L.E., Wooldridge, J.M., 2005. A computational trick for delta-method standard errors. Economics Letters 86 (3), 413–417.

Raven, J.C., 1982. Revised Manual for Raven's Progressive Matrices and Vocabulary Scale. NFER Nelson, Windsor, UK.
Shipley, W.C., 1946. Institute of Living Scale. Western Psychological Services, Los Angeles.
Thorndike, R.L., 1942. Two screening tests of verbal intelligence. Journal of Applied Psychology 26, 128–135.
Thorndike, E.L., Bregman, E.O., Cobb, M.V., Woodyard, E., 1927. The Measurement of Intelligence. Teachers College Bureau of Publications, New York, NY.
van der Linden, W.J., 2005. Linear Models for Optimal Test Design. Springer, New York.
Wechsler, D., 1981. Manual for the Wechsler Adult Intelligence Scale-Revised. The Psychological Corp., New York.
Zimowski, M.F., Muraki, E., Mislevy, R.J., Bock, R.D., 1996. BILOG-MG [Computer Software]. Scientific Software International, Chicago.