ED 462 403 TM 033 665

AUTHOR	Berk, Eric J. Vanden; Lohman, David F.; Cassata, Jennifer Coyne
TITLE	What Does a Verbal Test Measure? A New Approach to Understanding Sources of Item Difficulty.
PUB DATE	2001-04-00
NOTE	25p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*College Students; *Construct Validity; Higher Education; Item Response Theory; *Test Items; Test Results; *Verbal Tests
IDENTIFIERS	Self Report Measures

ABSTRACT

Assessing the construct relevance of mental test results continues to present many challenges, and it has proven to be particularly difficult to assess the construct relevance of verbal items. This study was conducted to gain a better understanding of the conceptual sources of verbal item difficulty using a unique approach that integrates self-report information with item response theory (IRT) data. Sources of difficulty were examined for sentence completion items on a standardized multiple-choice test, the Cognitive Abilities Test (R. Thorndike and E. Hagen, 1993) as part of the development of a new edition of the test. Results from a sample of 35 college students indicate that while there is a unidimensional nature to sentence completion solutions, specific sources of verbal features that can be manipulated easily to generate items for students at different levels of ability remain difficult to identify. This has implications for the construction and manipulation of construct-relevant variance in verbal item pools. Limitations of this study, implications for future research, and the role of integrated self-report and IRT data are discussed. Two appendixes contain tables of factor analysis results. (Contains 3 tables and 16 references.) (Author/SLD)

What does a Verbal Test Measure? A New Approach to Understanding Sources of Item Difficulty

Eric J. Vanden Berk and David F. Lohman

The University of Iowa

Jennifer Coyne Cassata

Fairfax County Public Schools

Paper Presented at the Annual Meeting of the

American Educational Research Association

Seattle, Washington 2001

Abstract

Assessing the construct relevance of mental test results continues to present many challenges. Within each cognitive domain, accurate models of how individuals solve item problems must be generated. This necessitates an understanding of what makes items difficult for students at varying levels of ability. Verbal items have proven to be particularly difficult to understand in these terms. This is due in part to the multiple strategies involved within a single item solution and the high relations among the interacting solution attributes within the entire verbal domain. The purpose of this study was to gain a better understanding of the conceptual sources of verbal item difficulty using a unique approach that integrates self-report information with IRT data. Specifically, sources of difficulty were examined for sentence completion items on a standardized multiple-choice test, the Cognitive Abilities Test (CogAT; Thorndike & Hagen, 1993), as part of the development of a new edition of the test. Results indicate that while there is a unidimensional nature to sentence completion solutions, specific sources of verbal features, that can be easily manipulated to generate items for students at different levels of ability, remain difficult to identify. This has implications for the construction and manipulation of construct-relevant variance in verbal item pools. Limitations of the current study, implications for future research, and the role of integrated self-report and IRT data are discussed.

Author Note

Correspondence concerning this article should be addressed to Eric J. Vanden Berk, the College of Education, University of Iowa, N440 Lindquist Center, Iowa City IA, 52242. E-mail may be sent to eric-vanden@uiowa.edu.

What does a Verbal Test Measure? A New Approach to Understanding Sources of Item Difficulty

Beginning in the 1970's, several teams of investigators studied cognitive tests as cognitive tasks. The main goal of this research was to understand how examinees of differing abilities solve items on the test. Models of how individuals solve items are grounded in an understanding of what makes items differ in difficulty. For example, Sternberg (1985) argued that intelligence tests measure metacomponential analytic functioning, including strategy selection, monitoring, and allocation of resources. In contrast, Carpenter, Just, and Shell (1990) argued that working memory capacity provides a better explanation for individual differences in performance on abstract reasoning tests. Embretson (1995) compared these two theories and found that general control processing was a more important aspect of the general intelligence measured by abstract reasoning tests than was working memory capacity.

Understanding what makes items difficult is thus an important first step in understanding how solutions might be processed. Such investigations have been much more successful in the figural domain (Carpenter et al., 1990; Embretson, 1995) than in the verbal domain. Indeed, several efforts to generate new items for complex verbal tasks (such as verbal analogies) from models of item difficulty have not succeeded (Bejar, Chaffin, & Embretson, 1991). This is not to say that we have not learned a lot about what makes verbal tasks difficult, and how these sources of difficulty can be manipulated to increase the construct-relevant (or irrelevant) variance in tests (Lohman, 2000). Rather, the claim is that we do not understand sources of difficulty in verbal tasks very well, and what we do understand often does not generalize well (Buck, Van Essen, Tatsuoka, Kostin, Lutz, & Phelps, 1998).

Within the verbal domain, many researchers have focused on understanding the relations among words and concepts from a variety of perspectives. Analytic approaches to studying these relations have led to the development of general taxonomies of semantic relations as well as taxonomies for verbal analogy items (Bejar et al., 1991). Bejar et al. (1991) initially developed a taxonomy of semantic relations for the analogy items used on the Graduate Record Examination (GRE). Taxonomies of this type could be useful tools in the construction of sets of verbal analogy items because systematic approaches can be followed and the various classes of relations can be represented.

Recently, the development of latent semantic analysis (LSA), in which mathematical approaches are used to extract the relations among words from passages, have been considered as models of knowledge acquisition (Landauer, Foltz, & Laham, 1998). The developers of LSA proposed two possible ways of conceptualizing it, as either a method for characterizing word meaning based on associations inferred from text, or as a model of how people acquire and use knowledge (Landauer et al., 1998). In terms of its relationship to test construction, LSA has been assessed as a model of the choices people make on multiple-choice subject matter tests and has been used to predict the comprehensibility of text (Foltz, Kintsch, & Landauer, 1998). Once its practical usefulness is more firmly established, LSA could theoretically be used in the construction of sentence-completion tests by providing information about the sentences to be included as items.

Attempts have been made to identify common attributes underlying performance in the verbal domain using rule space methodology (Buck et al., 1998). Specifically, Buck et al. examined the attributes underlying the critical reading, sentence completion, and analogies sections of the SAT. They treated sentence completion tests as tests of reading comprehension, coding four types of attributes for each item: vocabulary knowledge in both stem and options, syntactic complexity, rhetorical and semantic structure, and content. The rule space analysis retained a set of attributes and interactions that all correlated with total score on the test, indicating that many variables influence the processing involved in sentence-completion items (Buck et al., 1998; Lohman, 2000). In the construction of sentence-completion items, then, multiple variables can be manipulated to influence item difficulty. However, these variables tend to be highly related with one another and cannot be neatly disentangled.

It is these high relations among the interacting pieces within verbal items that make it more difficult to generate items at varying levels of difficulty in the same way as with mathematics items, when the many operations involved can be varied slightly to create vastly different levels of difficulty. In addition, it makes the problem-solving process more complicated for test-takers, because there are typically a variety of strategies for them to choose from when solving verbal items (Buck et al., 1998).

By understanding the different factors that make all types of items difficult for students to complete, especially on ability tests such as the CogAT, those factors can be manipulated and items can be generated at varying levels of difficulty. With multi-level tests such as the CogAT, which is used with students from third to twelfth grade, it is critical to generate items at all points on the difficulty continuum, since some test items overlap with adjacent grade levels. However, building verbal test items from existing taxonomies of verbal item functioning continues to present many challenges.

Sheehan and Mislevy (2001) recently studied the nature of sentence completion tasks by investigating item features and expert test developer's difficulty estimates. In addition, they examined certain item variants that could be isolated. In comparison to IRT estimates, the judges were found to be proficient at estimating difficulty for "definitional" items but not for "reasoning" items. No significant effects were found for item features such as sentence length, negation and sentence structure. However, item variant investigations revealed some important findings. For instance, adding difficult vocabulary words to a sentence does not seem to alter item difficulty unless the word is necessary to convey the intended semantic relation. Also, substituting words that change vocabulary demand can substantially vary item difficulty.

As Sheehan and Mislevy's (2001) findings illustrate, item difficulty can be approached from the performance end. With the advent of newer and more sophisticated models of item calibration, the number of ways in which the construction of verbal items can be informed has dramatically increased. Given its salience over the past decade, item response theory (IRT) is clearly the most notable of these methods. Though it is typically used to assess the reliability of test instrument results (Camilli, 1999; Hambleton & Swaminathan, 1985), the use of IRT as a tool for investigating the construct validity of tests has gained some support. McNamera (1990), for example, used IRT to assess the construct validity of a language test within the Occupation English Test. In this study he illustrates how IRT can be used to construct a single dimension using data from a two parts of a test. Moreover, he found that ability estimates could provide significant information about the level of ability required to successfully solve different language tasks. Perhaps most importantly, the study demonstrated how a measurement model, working in harmony with models of language skill and ability that underlie test performance, can produce more meaningful sets of items.

Objections to the use of IRT typically concern violations of the unidimensionality and local independence assumptions (Hambleton, 1990). That is, most IRT models specify that the items of a single test measure a single ability. Since most all item responses are multiply determined, this is a difficult requirement to satisfy--often justifying the objections. The unidimensionality requirement is even more problematic for verbal tests, since responses to any single verbal item tend to have more response factors than found in most other domains. Thus, the practicality of IRT to assess any aspect of a verbal test instrument, much less construct validity, is largely dependent on satisfying the unidimensionality--and by extension, local independence--assumption. It is important to also note that although multi-dimensional data sets can be constructed and accurately calibrated, standardized ability tests overwhelmingly assume unidimensionality.

If, however, a data set is shown to be sufficiently unidimensional, we can identify the contribution of any single item to test validity independent of the other items. In this way IRT models can potentially be used to compliment other sources of information about the items and construct models. This includes theoretical models that reflect our most current thinking about the cognitive processes and abilities that underlie verbal task performance. Since these cognitive models are independent and cannot be constructed solely from measurement models, test developers must look to other sources of item information.

Self-report information is often used for its capacity to gain a richer understanding of certain conceptual sources of difficulty that may underlie task performance. While self-report data are useful for constructing models of underlying test constructs, they are often unreliable and can be of questionable accuracy. Since IRT calibrations can provide much more reliable item parameter estimates, it is possible that the two sources of data can compliment each other in such a way that it creates a more informed understanding of the sources of item difficulty.

The present study was designed to investigate a unique approach to understanding what makes verbal items difficult, what items at different difficulty levels have in common, and how students of differing abilities approach these items. Sentence completion items from the newly standardized Cognitive Abilities Test (CogAT) were used. Sentence completion items are particularly interesting because they represent the oldest verbal mental test item format, still are widely used, yet remain virtually unstudied (Buck et al., 1998; Just & Carpenter, 1987). By complimenting self-report information with accurate IRT models, it is possible that a better understanding of the conceptual sources of item difficulty will emerge. Accurate taxonomies of verbal item functioning depend on such research and are not only used to build cognitive models but are of crucial importance to test developers interested in establishing evidence of construct validity.

Method

Two distinct approaches were integrated in the present study. The first approach involved collecting self-report information. While there is sufficient evidence to expect that test takers are willing to perform self-report tasks, it is unclear whether or not test takers have the capacity to articulate the conceptual sources of item difficulty on sentence completion items. For this reason, a sample of college students (n = 35) at a large mid-western university were asked to complete the sentence completion tests intended for grade twelve and provide self-reports.

Students were first asked to complete both the sentence completion and verbal analogies sub-tests of the highest level (intended for grade twelve) of the new edition of the CogAT. Next, students were asked to complete a questionnaire about how they approached the sentence completion items. In the questionnaire, students were first asked to list the items that they found particularly easy and difficult. Then, students responded to questions about the strategies that they used to solve both types of items. Student responses were grouped into categories and analyzed for the whole group and across score levels.

Using BILOG, item response theory (IRT) parameter estimates were then obtained for all grades on the national tryout data. However, for such IRT results to be of use, it was essential to first establish the unidimensionality of the data sets. For this reason, two independent methods of dimensionality assessment were incorporated in addition to the IRT fit statistics. First, student scores on two separate forms of the sentence-completion items from the national tryout sample were assessed for dimensionality using Bock's full-information factor analysis routine (Bock, Gibbons, & Muraki, 1988) provided by TESTFACT (see Wilson, et. al, 1984). These same data were then assessed for essential-unidimensionality using DIMTEST (see Stout et. al,1993). See appendix A for a summary of the full-information factor analysis results. See appendix B for a summary of the DIMTEST essential-unidimensionality results.

Once essential unidimensionality was established for all grades it was possible to use IRT to place the items on a common scale. Items were then ranked in order to compare item difficulty to the self-report information for the common items. As a further step, an IRT analysis was run with the standardization sample at all grades (from three to twelve). Again, items were ranked in terms of difficulty and compared to the analyses of the grade twelve items from the national tryout sample. Here, we attempted to discover common factors that may be contributing to item difficulty. In addition, students' ratings of items as easy and/or difficult on the self-report questionnaires were compared to the actual difficulty levels.

Results

The present study summarizes both what college students reported about their solution strategies for sentence-completion items and how this information relates to IRT analyses of item scores from an administration of the same items to national standardization samples of high school students. The self-report data from the college student sample are summarized first. It is important to mention that these students, as expected, performed quite well on the sentence-completion items, with 26 of the 35 students answering at least 14 of the 20 items correctly (mean = 14.86, SD = 3.01). Patterns were found at different score levels, however, in terms of the complexity of the strategies students mentioned for approaching items. When asked, "how did you go about figuring out the correct answers for the sentence completion items?" students listed a variety of strategies. Among students with the highest scores, the most frequently mentioned strategies included, "look for relationships among words or meanings," and "look for a match with my own knowledge or ideas." In contrast, as you moved down the score scale, the most common strategies included, "plug each word into the sentence;" "pick what sounds best," and "process of elimination."

In general, student responses were quite varied about whether they used the same strategies for both easy and difficult items. Eighteen students responded that they did use the same strategies, and 13 students responded that they used different strategies. No discernible differences were found across score levels. Students at all score levels were less sophisticated when describing specific strategies for either the easy or difficult items. In response to the question about the more difficult items, the strategies mentioned tended to be simpler and were similar across score levels, such as, "plug each option into the sentence," or "process of elimination." With respect to specific items, students mentioned many different items as being particularly easy or difficult. The following section contains a discussion of the results of IRT analyses with both the national tryout and standardization samples.

IRT Results for Standardization and National Tryout Samples

Fit statistics indicated that the three-parameter logistic model (PLM) was most appropriate for the national tryout samples. On form B grade 12, using a 1 PLM, 14 of 24 items had χ^2 fit p-values above .05; 20 of 24 items fit using the 2 PLM; and 23 of 24 items fit using a 3 PLM. Similar results were found for form A grade 12, with all 24 items displaying χ^2 fit p-values above .05 when using a 3PLM. Results consistently supported the 3PLM at each grade level as well. This was expected since a model that accounts for discrimination and the effects of guessing is typical of well-written multiple-choice items designed for norm-referenced purposes.

Since the standardization sample consisted of a sample size in excess of 150,000 examinees, χ^2 fit statistics--notorious for overestimating models when sample sizes.are large--were not a useful guide to model fit for these data. Given the reasonable evidence for the 3PLM solution in the national tryout results, coupled with the dimensionality assessment results, it was concluded that the 3PLM was the most appropriate solution for the standardization sample data as well.

IRT parameter estimate correlations between the standardization sample and the national tryout sample were quite high (difficulty r=.9339; discrimination r=.8421). There was also a good range of difficulty estimates, where item 7A was the least difficult (national tryout: *b*=1.846; standardization: *b*=.6163), and item 24B was the most difficult (national tryout: *b*=-2.7018; standardization: *b*=-2.2267). Again, this is expected of well-written items designed for norm-referenced purposes. Finally, discrimination estimates were all sufficiently high and positive (national tryout: mean = 2.1109, SD = .7142; standardization: mean = 1.1456, SD = .3251). See table 1 for a full description of all item parameters for the items appearing in the self-report study.

Initial IRT analyses conducted with the national tryout data found essential unidimensionality for all grade levels of the CogAT that were examined. This has important implications. For one, it demonstrates that the data are robust with respect to unidimensionality and local independence assumptions. Thus, we have some assurance that our item parameter estimates are reliable. Moreover, we can be reasonably certain that our measurement model is stable enough to be used in tandem with the model(s) formed by the self-report data and the theoretical models of the constructs believed to underlie cognitive solutions in the verbal task domain.

On the other hand, one might find it contradictory that the latent space underlying sentence completion items contains only a single dimension while self-report data, and theoretical models suggest that students use a diversity of strategies. However, closer analysis and understanding suggest that the two are not mutually exclusive. Although the test measures a single dominant underlying construct, the solution strategies employed by students across ability levels can be quite diverse in nature. It is possible that these various strategies may be influenced by some common set of item or task characteristics that contribute to their difficulty. Though our measurement model cannot answer this, unidimensionality findings do suggest further exploration of this possibility.

Comparison of Self-Report Data with IRT Data

For each data set, the items were ranked in terms of difficulty and these ranks were correlated using the Spearman-rho correlation coefficient. The correlations between the self-report ranks and the two other ranks were significant and quite high ($\underline{r}$ = .798 for self-report and tryout and $\underline{r}$ = .686 for self-report and standardization). When we ran a regression using Self-Report Rank as our independent variable and the Standardization and National Tryout samples as or dependent variables we obtained an R-square of .69 and significance at the .01 level.

With respect to specific items, students mentioned many different items as being either particularly easy or difficult. Student ratings were compared to the difficulty rankings for the items gathered from the IRT analyses of the standardization data. With only one exception, student descriptions of items as easy corresponded to their actual difficulty. One item that was relatively difficult was described as easy by one student. A similar pattern was found for student descriptions of items as difficult. More items were described as being difficult than as being easy and the ratings were more dispersed. One possible explanation for this consistency is that the students who participated in the self-report study all tended to perform well on the test, which could have made it easier for them to accurately reflect on their performance.

Characteristics of Items

Once the items were ranked and there was assurance that the self-report difficulty rankings were consistent with actual item difficulty ratings, several characteristics of the items were examined in an attempt to gather information about the features that may have contributed to item difficulty. The number of words in the sentences was not found to relate to self-report difficulty ratings ($\underline{r}$ = -.03), tryout ($\underline{r}$ = .12) or standardization difficulty ratings ($\underline{r}$ = .06), indicating that length of sentence did not contribute to item difficulty. Perhaps, with the full set of items representing the larger span of difficulty, the number of words may be a significant factor in item difficulty. It is also possible that sentence length may have a positive association with informational cues about the correct answer.

Six of the 20 items used in the self-report study and standardization each had a missing adjective, noun, or verb, respectively with one item missing a conjunction and the other an adverb. Items focused on these various parts of speech were distributed throughout the levels of difficulty, with adjective items occurring slightly more frequently at the easier end of the difficulty scale (three of the five easiest items). Again, with the full set of items representing the larger span of word types, the parts of speech may be a significant factor in item difficulty.

Discussion

IRT analyses of the national tryout and standardization samples of the CogAT sentence completion items were compared to self-report information for 20 of these items in order to investigate the sources of item difficulty. For the IRT analyses to be of use, it was first necessary to carefully establish the unidimensionality of the data sets. The independent findings for a single dimension for items on the sentence completion portion of the test confirm that it is possible to isolate a verbal ability model and that verbal ability can be measured as a cognitive task. Furthermore, it suggest that sources of difficulty, though varied, may have an identifiably commonality. With the unidimensionality assumption satisfied, we can be reasonably certain that any single item's contribution to validity does not depend on what other items were included in the test. Thus for any item in the self-report study, a more powerful item calibration model was compared.

Despite being the oldest verbal mental test item format and their widespread use, very little is known about how sentence completion items function. Thus, there is no single best criterion upon which to base item selection. While it would be tempting to select items primarily on robust IRT calibrations, our self-report information suggest that even the best measurement model is not enough to describe verbal item difficulty. Previous research in the verbal domain suggests that general taxonomies of verbal functioning can be produced for particular groups of items but do not generalize well. This was confirmed in our study, which suggests the limitation of such taxonomies as useful tools for generating large sets of verbal items along the difficulty continuum.

The present study indicates, in a preliminary way, that combining subjective information with IRT analyses could help in the development of verbal test items. College students were used since the capacity of test takers to describe item difficulty was unknown. The appropriate next-step is to conduct self-reports with test takers of different ages using grade-appropriate tasks. It may also be beneficial to gather the self-report information on-line and item-by-item. If this approach were used, it would be critical to develop appropriate procedures for gathering self-report information, as the questions used in the present study did not yield very specific information. Early in the item development process, students who resemble potential examinees could be interviewed about their approaches to particular items and engaged in discussions about what makes the items difficult. This approach provides useful validity evidence for the difficulty of items, rather than relying solely on measurement and statistical models.

Results from the present study illustrate the complexity involved in generalizing about what makes verbal items difficult. Self-report data from examinees can provide useful information for item development when used in conjunction with objective measures of item difficulty. However, this study also demonstrates that information about item difficulty does not guarantee an understanding of the specific sources of that difficulty. Unlike with items in other areas, such as mathematics and spatial items, it is quite difficult to determine which tangible characteristics of items contribute to item difficulty. As a result, it is quite a challenge to develop a blueprint for how verbal items, and especially sentence-completion items, can be generated. It appears that many of the sources of difficulty are either intangible or result from the interaction of many factors, some within the items and some related to the cognitive skills of examinees.

As a result, to gather information about these difficulty factors, it is necessary to examine not only the items themselves, but also to engage examinees in discussions about those items. Past studies of sentence-completion items have treated them as measures of reading comprehension (Buck et al., 1998). However, since each item is limited in length, it is apparent that the comprehension and understanding of the words with little context requires additional knowledge and reasoning skills. Studies with actual examinees would provide useful validity evidence about these skills that could not be gathered solely through statistical and measurement techniques.

References

Bejar, I., Chaffin, R., & Embretson, S. (1991). Cognitive and psychometric analysis of analogical problem solving. NewYork: Springer-Verlag.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information factor analysis. Applied Psychological Measurement, *12(3),* 261-280.

Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., & Phelps, M. (1998). Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I Verbal: Sentence Completion section. (Research Rep. No. 98-23). Princeton, NJ: Educational Testing Service.

Camilli, G. (1999). Measurement error, multidimensionality, and scale shrinkage: A reply to Yen and Burket. Journal of Educational Measurement, 36, 73-78.

Carpenter, P.A., Just, M.A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. Psychological Review, 97, 404-431.

Embretson, S.E. (1995). The role of working memory capacity and general control processes in intelligence. Intelligence, 20, 169-189.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nyijhoff.

Just, M.A., & Carpenter, P.A. (1987). The psychology of reading and langauge comprehension. Newton, MA: Allyn and Bacon.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25, 259-284.

Lohman, D.F. (2000). Intelligence and complex information processing. In R.J. Sternberg (Ed.), Handbook of intelligence (pp. 285-340). Cambridge, UK: Cambridge University Press.

McNamara, T.F. (1991). The role of item response theory in language test validation. In S. Anivan (Ed.) Current Development in Language Testing. Singapore: SEAMEO Regional Language Center.

Sheehan, K. M. & Mislevy, R. J. (2001). An Inquiry into the Nature of the Sentence Completion Task: Implications for Item Generation. (ETS Research Report). Princton, NJ: Educational Testing Service.

Sternberg, R.J. (1985). Beyond IQ: A triarchic theory of human intelligence. Cambridge, UK: Cambridge University Press.

Stout, W. F., Douglas, J., Junker, B., & Roussos, L. (1993). DIMTEST manual. Champaign, IL: University of Illinois at Champaign-Urbana.

Thorndike, R.L., & Hagen, E.P. (1993). Form 5 CogAT research handbook: All levels. Chicago: Riverside.

Wilson, D, Wood, R. L. & Gibbons, R. (1984) TESTFACT: Test scoring and item factor analysis. Chicago: Scientific Software, Inc.

Table 1 IRT item parameter estimates for all items in the self-report study

Form A		3PLM				
Item		Difficulty	Discrimination	Chi-Square Fit (p-value)	Rank by diff.	Self-Report (Part of Speech)
3	tryout	1.655713	2.240958	(0.9807)	2	2 people labeled easy (Noun)
	stand.	0.091293	1.456020	(0.0000)	3	
	SR				1	
5	tryout	0.627239	1.175341	(0.1222)	6	1 person labeled easy (Verb)
	stand.	-0.368540	1.145393	(0.0010)	7	
	SR				9	
7	tryout	1.846762	2.525806	(0.8509)	1	7 people labeled easy – easy b/c wording cues, common language (Adverb)
	stand.	0.616255	1.687495	(0.0001)	1	
	SR				1	
8	tryout	0.712642	2.667469	(0.7024)	5	2 people labeled easy 1 person labeled hard (Adjective)
	stand.	-0.437732	1.575323	(0.0000)	9	
	SR				1	
9	tryout	0.612715	1.396963	(0.8905)	8	4 people labeled easy - easy b/c wording (Adjective)
	stand.	-0.122215	0.761548	(0.0012)	4	
	SR				5	
17	tryout	0.666193	2.427345	(0.8939)	7	1 person labeled easy 3 people labeled hard (Verb)
	stand.	-1.209788	1.201401	(0.0049)	15	
	SR				4	
18	tryout	-0.609278	2.535446	(0.4933)	15	2 people labeled hard (Verb)
	stand.	-1.426335	1.333552	(0.5579)	16	
	SR				11	
23	tryout	-0.380241	1.941757	(0.1013)	14	3 people labeled hard -b/c people said there were multiple possible correct answers (Adjective)
	stand.	-0.811594	0.863712	(0.1647)	11	
	SR				19	
24	tryout	-1.337777	1.816976	(0.6278)	17	1 person labeled hard (Noun)
	stand.	-1.702183	0.919661	(0.1131)	17	
	SR				18	

National Tryout (tryout) N=398; Standardization (stand.) N=150,000; Self Report (SR) N=35

Form B		3PLM				
Item		Difficulty	Discrimination	Chi-Square Fit (p-value)	Rank by diff.	Self-Report (Part of Speech)
1	tryout	1.289333	1.438861	(0.0068)	3	3 people labeled easy 3 people labeled hard (Adjective)
	stand.	0.196633	0.909152	(0.0023)	2	
	SR				5	
5	tryout	0.479994	1.913125	(0.0917)	9	1 person labeled hard – hard b/c cliché (Noun)
	stand.	-0.596810	1.114348	(0.0019)	10	
	SR				10	
12	tryout	0.938147	1.890947	(0.2247)	4	2 people labeled easy (Adjective)
	stand.	-0.214626	1.014687	(0.0356)	5	
	SR				5	
16	tryout	-1.815863	3.393287	(0.5320)	19	SR diff rank = 13 (Adjective)
	stand.	-2.052761	1.691645	(0.1609)	19	
	SR				13	
17	tryout	0.405474	1.485549	(0.0949)	10	1 person labeled easy 3 people labeled hard (Verb)
	stand.	-0.386672	0.874009	(0.0053)	8	
	SR				13	
18	tryout	-1.411510	3.883987	(0.7850)	18	SR diff rank = 15 (Verb)
	stand.	-1.662275	1.690031	(0.1297)	18	
	SR				15	
19	tryout	-0.291524	1.201357	(0.4822)	12	1 person labeled hard (Conjunction)
	stand.	-1.094435	0.714664	(0.0024)	14	
	SR				16	
21	tryout	-0.743351	2.658186	(0.1082)	16	1 person labeled hard (Noun)
	stand.	-0.971725	1.168634	(0.5579)	12	
	SR				5	
22	tryout	0.055511	2.055081	(0.8508)	11	1 person labeled hard (Noun)
	stand.	-0.292655	0.969187	(0.0197)	6	
	SR				11	
23	tryout	-0.314920	1.366057	(0.5115)	13	2 people labeled hard (Verb)
	stand	-0.974514	0.758685	(0.0341)	13	
	SR				16	
24	tryout	-2.701821	2.204257	(0.2488)	20	4 people labeled hard (Noun)
	stand.	-2.226723	1.062022	(0.1012)	20	
	SR				20	

National Tryout (tryout) N=395; Standardization (stand.) N=150,000; Self Report (SR) N=35

Appendix A
Full-information Factor Analysis of CogAT Verbal Sentence Completion Form A

Grade 4

Number of factors in the solution	Latent Root	% Variance Explained	χ^2	Df	χ^2 Change	Df Change	$P(\chi^2)$ Change
1	10.51	34.02	5409.5*	445			
2	1.65	7.24	5290.6*	422	118.85*	23	0.00
3	0.86	3.80	5293.4*	400	LNI	-	-
4	0.65	2.79	5265.7*	379	27.65	21	0.15

C=7,8,6,6,9,6,11,7,9,16,10,19,11,8,7,8,8,11,4,19,5,6,6,21

Grade 6

Number of factors in the solution	Latent Root	% Variance Explained	χ^2	Df	χ^2 Change	Df Change	$P(\chi^2)$ Change
1	4.59	20.74	2920.5*	401			
2	1.51	8.69	2899.2*	378	21.40	23	0.56
3	1.29	4.73	2878.1*	356	20.85	22	0.53
4	1.13	4.48	2920.0*	335	LNI	-	-

C=10,13,9,9,14,8,8,6,12,12,6,9,11,8,16,12,8,10,24,12,11,6,12,19

Grade 8

Number of factors in the solution	Latent Root	% Variance Explained	χ^2	Df	χ^2 Change	Df Change	$P(\chi^2)$ Change
1	10.68	36.78	5382.6*	344			
2	0.99	4.00	5379.1*	311	3.41	23	1.00
3	0.89	3.92	5410.1*	289	LNI	-	-
4	0.76	2.78	5335.8*	268	74.37	24	0.00

C=12,14,12,13,11,9,11,13,13,17,23,15,13,9,15,16,17,9,13,12,22,22,12,15

Grade 10

Number of factors in the solution	Latent Root	% Variance Explained	χ^2	Df	χ^2 Change	Df Change	$P(\chi^2)$ Change
1	11.0	38.8	6399.0*	386			
2	0.99	4.26	6406.6*	363	LNI	-	-
3	0.90	3.47	6414.3*	341	LNI	-	-
4	0.72	2.20	6338.6*	320	75.64	21	0.00

C=14,12,18,12,16,15,17,14,11,9,15,17,20,20,23,25,21,26,32,13,12,5,18,15

Grade 12

Number of factors in the solution	Latent Root	% Variance Explained	χ^2	Df	χ^2 Change	Df Change	$P(\chi^2)$ Change
1	10.9	33.17	5644.6*	327			
2	1.11	4.39	5619.8*	304	24.78	23	.36
3	0.73	3.55	5652.1*	282	LNI	-	-
4	0.70	3.19	5610.9*	261	41.13	21	.005

C=9,12,12,9,11,10,13,17,13,13,12,15,12,12,13,12,14,9,14,14,14,17,5,9

C=each items respective guessing parameter as estimated using BILOG
*Significant at the p.<.001 level
Note: Latent roots and variances explained are from the 4-factor solutions.
LNI= Likelihood does not increase.

Appendix B
DIMTEST Analysis of CogAT Verbal Completion Items
Form A

Grade 4		N=523	n items=24
Conservative T		More Powerful T	
T	P	T	P
-1.709	.9563	-1.7429	.9593

AT1: 7 23 19 17 21 24
AT2: 22 16 18 15 13 4

Grade 6		N=633	n items=24
Conservative T		More Powerful T	
T	P	T	P
.4727	.3182	.7656	.222

AT1: 22 21 15 20 24 23
AT2: 19 11 18 14 13 17

Grade 8		N=406	n items=24
Conservative T		More Powerful T	
T	P	T	P
-.1393	.5554	.0803	.5320

AT1: 4 9 10 6 3 16
AT2: 1 2 15 17 19 23

Grade 10		N=470	n items=24
Conservative T		More Powerful T	
T	P	T	P
-0668	.5266	-.0256	.5102

AT1: 17 15 24 4 21 5
AT2: 1 3 19 16 10 23

Grade 12		N=398	n items=24
Conservative T		More Powerful T	
T	P	T	P
-1.3945	.9185	-1.6456	.9501

AT1: 9 12 11 7 4 1
AT2: 3 15 6 8 16 21

H_o: d_E=1 H_1:d_E>1

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM033665

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: What does a Verbal Test Measure? Understanding Sources of Verbal Item Difficulty

Author(s): Eric J. Vanden Berk; David F. Lohman; and Jennifer Coyne Cassata

Corporate Source: University of Iowa

Publication Date: 4/2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
↑ ☑	↑ □	↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:

Printed Name/Position/Title: Eric J. Vanden Berk/Student/PhD Measurement and Sta

Organization/Address:
University of Iowa
College of Education
N440 Lindquist Center
Iowa City IA 52242

Telephone: 319-335-6153

Fax:

E-mail Address: eric-vanden@uiowa.edu

Date: 2/2/02

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory (Bldg 075)
College Park, Maryland 20742

Telephone: 301-405-7449
Toll Free: 800-464-3742
Fax: 301-405-8134
ericae@ericae.net
http://ericae.net