An Analysis of Similarities in Item Functioning Within Antonym and Analogy Variant Families

David J. Scrams
Robert J. Mislevy
Kathleen M. Sheehan

June 2002
GRE Board Professional Report No. 95-17aP
ETS Research Report 02-13
An Analysis of Similarities in Item Functioning Within Antonym and Analogy Variant Families

David J. Scrams
Microsoft, Inc.

Robert J. Mislevy
University of Maryland

Kathleen M. Sheehan
Educational Testing Service

GRE Board Report No. 95-17aP

June 2002

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board and Educational Testing Service.

Educational Testing Service, Princeton, NJ 08541
Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do not necessarily represent official Graduate Record Examinations Board position or policy.

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logos, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service.

Educational Testing Service
Princeton, NJ 08541

Copyright © 2002 by Educational Testing Service. All rights reserved.
Abstract

This study examined the potential advantages of using item variants — sets of related test items based on a common template or schema — to offset the increased item-production demands associated with continuous computer-based testing in high-stakes assessments. Empirical results showed that the calibrated difficulty parameters of analogy and antonym item variants that had been used in operational administrations of the Graduate Record Examinations (GRE®) General Test were very similar. Results of a simulation study that considered precision losses within a linear testing framework projected precision losses of less than 10% for estimating examinee abilities from responses to variants when using expected response functions based just on variant-family information, compared to using true item response theory parameters for individual items.

Key words: Computerized adaptive testing, expected response function, item calibration, item response theory, item variants
Acknowledgements

This work was completed while David Scrams was a postdoctoral fellow at Educational Testing Service (ETS®) and Robert Mislevy was a member of the ETS research staff. Support was provided by the Graduate Record Examinations Research Committee. The authors are grateful to Adam Banta, Lily Knezevich, and Robin Jacobs for their assistance in obtaining the data, and to Marilyn Wingersky for providing the expected-response-function solutions used in the simulations.
Table of Contents

Introduction......................................................................................................................................1
Data..................................................................................................................................................2
Results..............................................................................................................................................4
    Empirical Analyses ..............................................................................................................4
    Simulations ...........................................................................................................................7
Discussion......................................................................................................................................11
References......................................................................................................................................14

List of Tables

Table 1. Word Substitutions Used to Produce Item Variants ......................................................3
Table 2. Covariance Matrices from Multivariate Multiple Regression Analyses .......................6

List of Figures

Figure 1. Scatterplots of Variant Item Parameters Against Parent Item Parameters for Analogies and Antonyms ........................................................................................................5
Figure 2. Mean Estimated Theta Using True, Multivariate Multiple Regression Analysis and Expected-Response-Function Item Parameters for Analogy and AntonymVariant Simulations..........................................................................................................................10
Figure 3. Root Mean Squared Error Using True, Multivariate Multiple Regression Analysis and Expected-Response-Function Item Parameters for Analogy and AntonymVariant Simulations..................................................................................................................10
Introduction

Computerized adaptive tests (CATs) offer a number of significant advantages over conventional paper-and-pencil administrations. Among these advantages are increased efficiency, continuous test availability, immediate score reporting, and the potential for broader ranges of item types (Green, 1983). But CATs are not without costs. The advantages mentioned can only be realized with rich item pools, so it is usually necessary to produce considerably more items for CATs than for conventional tests (Flaugher, 1990).

The distribution of difficulties needed for conventional tests and CATs differs as well. Conventional tests generally consist primarily of middle-difficulty items; they have fewer low-difficulty and high-difficulty items, because there are fewer examinees in those proficiency ranges. CATs, on the other hand, require sufficient numbers of items across the entire difficulty range. This need for a broader distribution of difficulty will further tax item production because easy and difficult items are harder to write than middle-difficulty items.

Satisfying content requirements across difficulty levels may also be problematic. And these richer CAT item pools may need to be replenished more frequently when a high-stakes testing program offers continuously available testing. All of the above factors can combine to dramatically increase item-production demands.

Producing item variants is an attractive response to increased item-production demands. Variants are items created from a common schema. Hively, Patterson, and Page (1968) and Embretson (1998), for example, show how item variants can be generated from theoretically derived templates. Less formally, test developers can write variations of items around the same essential elements. Or they can formulate several variations from a single originating item — that is, several “children” from the same “parent” item. Creating families of related items may require less effort than creating the same number of items independently.

Increasing item production is clearly a necessity for programs that move from paper-and-pencil to CAT administration, but increased production alone is not enough. New items must be calibrated onto existing item response theory (IRT) scales before being used operationally. The standard method of calibrating items is through item pretesting. New items are administered to a large sample of examinees, along with some items that had previously been calibrated into the scale, and the parameters of the new items are then estimated on the basis of examinee responses to both the new and old items. Thus, increased item production causes increased item-pretesting
demands. However, the desire for shorter test lengths (to realize CAT’s promise of increased test efficiency) can produce a bottleneck for item pretesting. Standard item calibration approaches, with typical calibration sample sizes and greater numbers of new items, could require each examinee to answer more pretest items than are necessary under conventional testing.

The current study focuses on another potential benefit of item variants that addresses the problem of increased item-pretesting demands. Variant family members may have related, and therefore partially predictable, item parameters. With this additional information about item parameters, fewer pretest examinees might be required before items can be used operationally (Mislevy, Sheehan, & Wingersky, 1993).

Differences in item functioning among variants that were generated formally from the same theoretical framework can often be modeled in terms of particular item manipulations (e.g., Enright, Morley, & Sheehan, 2002). In this study we considered informally generated families of variants under the weaker hypothesis that variants within such a family will merely function similarly. Such regularities might allow variants to be used operationally with usefully reduced pretesting sample sizes, even when they have not been created with a goal of either keeping their parameters constant or of manipulating them systematically. Our focus, then, was on methodology rather than item generation. We investigated this possibility through empirical analyses of a convenience sample of items variants used operationally on the Graduate Record Examinations (GRE®) General Test.

Data

The convenience sample consisted of data for analogy and antonym item variants that had been used on the verbal section of the GRE General Test. The data available included large-sample IRT calibrations performed using LOGIST (Wingersky, 1983; Wingersky, Barton, & Lord, 1982). All items had been calibrated according to the three-parameter logistic (3PL) IRT model using responses from approximately 1,000 examinees. The items had been presented along with other pretest items during a paper-and-pencil administration. Different subsets of items had been administered to different subsets of examinees, so no examinee saw all pretest items.

The sample contained a total of 26 analogy variant families and 36 antonym variant families. These items had not been systematically designed for the present analysis, but had been
written during the course of test developers’ operational work. Therefore, we did not know, for example, whether the members of the variant families had been written as children from an originating parent or whether they had been produced as variations on a common theme.

Since the same basic statistical approach would be applied in either case, we analyzed the data as if all variant families had arisen as parents and children. We used a tracking code that had been assigned to each item to arbitrarily designate a single item in each variant family as a parent and all other members of the family as its children. Altogether, then, these analyses addressed 73 analogy items (organized as 26 parents and 47 children) and 102 antonym items (36 parents and 66 children). Fourteen of the 26 analogy families and 23 of the 36 antonym families included only two items (one parent and one child). The largest analogy family included seven items, while the largest antonym family included 10 items.

Since the variants had been produced following standard item-writing procedures, many different types of manipulations had been used in their production. Table 1 illustrates the simplest of these manipulations — individual word substitutions — using a pair of antonym variants. More complex manipulations included cases in which only the deep structure of variants remained intact (e.g., analogy-variant families in which all members were developed to conform to a common rationale). All of the variants had been produced to serve the goal of generating items efficiently without regard to either preserving or systematically influencing item parameters.

Table 1

Word Substitutions Used to Produce Item Variants

<table>
<thead>
<tr>
<th>Antonym family X, variant 1</th>
<th>Antonym family X, variant 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROPENSITY</td>
<td>PROPENSITY</td>
</tr>
<tr>
<td>*(A) antipathy</td>
<td>*(A) aversion</td>
</tr>
<tr>
<td>*(B) violation</td>
<td>*(B) violation</td>
</tr>
<tr>
<td>*(C) competence</td>
<td>*(C) competence</td>
</tr>
<tr>
<td>*(D) independence</td>
<td>*(D) independence</td>
</tr>
<tr>
<td>*(E) penalty</td>
<td>*(E) disadvantage</td>
</tr>
</tbody>
</table>
Results

Our discussion of results is divided into two sections: Empirical Analyses and Simulations. Under Empirical Analyses, we describe the analyses and results used to evaluate the item-functioning relationships among variants within families. In the Simulations section, we consider the effect on ability estimation within a linear testing format when variants are administered without pretest data that is specific to them; we do this using only information available from previously-calibrated members of the same variant families. These simulations are not meant to justify using variants without pretesting, but to focus on the utility of intrafamily relationships in a way that is easy to interpret.

Empirical Analyses

Figure 1 presents bivariate scatterplots that depict the relationships among IRT parameters of item variants; these relationships are shown as plots of child item parameters against their parent item parameters. Statistically reliable positive correlations between child and parent difficulty resulted for both analogies ($r = .56$) and antonyms ($r = .73$). Also, a reliable positive correlation between variant and parent discrimination resulted for antonyms ($r = .43$). The remaining correlations were positive, but not significantly different from zero.

For each item-type data set, a multivariate multiple regression (MMR) analysis was used to model the estimated IRT parameters of all available variant children as a function of the estimated IRT parameters of their respective parents. Each child variant was treated as an independent case in the analysis. Three dependent variables were included for each child; these corresponded to its own 3PL parameters ($a$, $b$, and $c$, or discrimination, difficulty, and asymptote, respectively). Three independent variables were also included for each child; these corresponded to the 3PL parameters of its parent ($a$, $b$, and $c$). If the child variants had been produced by systematically manipulating specific parent features, sets of independent variables indicating the specific types of manipulations implemented could also have been included in the analyses.
Figure 1. Scatterplots of variant item parameters against parent item parameters for analogies and antonyms.
Table 2 presents the resulting total and residual covariance matrices. The total covariance matrices provide an indication of overall item-parameter variability (shaded cells) and variability among item parameters within variant families (nonshaded cells). The residual covariance matrices indicate variability beyond that which is predictable from parent item parameters. For example, if variants within a family always had item parameters identical to those of their parents, the residual covariance would be a matrix of zeros. If the parameters of variants within a family were no more strongly related to one another than to items of any other family, the residual covariance matrix would be only slightly reduced from the total covariance matrix.

Table 2

<table>
<thead>
<tr>
<th>Covariance Matrices from Multivariate Multiple Regression Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariance matrix</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td><strong>Analogy variant children</strong></td>
</tr>
<tr>
<td>a</td>
</tr>
<tr>
<td>b</td>
</tr>
<tr>
<td>c</td>
</tr>
<tr>
<td><strong>Antonym variant children</strong></td>
</tr>
<tr>
<td>a</td>
</tr>
<tr>
<td>b</td>
</tr>
<tr>
<td>c</td>
</tr>
</tbody>
</table>

As Table 2 indicates, parent item parameters provided the most information about difficulty \( (b) \), explaining approximately 33% of the variability for analogy variants (which drops from 1.65 to 1.11) and 55% of the variability for antonym variants (which drops from 1.44 to 0.65). A fair amount of variability in item discrimination \( (a) \) was explained for antonym variants (approximately 21%), but little information was provided about item discrimination \( (a) \) for analogies (7%) or about the lower-asymptote parameter \( (c) \): 2% and 11% for analogies and antonyms, respectively. These analyses were conducted using discrimination and lower-
asymptote parameters in their standard scale. The analyses were repeated for both analogies and antonyms using the \( \log(a) \) and \( \text{logit}(c) \) parameters, but the results were comparable and are not reported here.

**Simulations**

Although the results from the multivariate analyses of variance are promising, they provide little direct information about the operational gains afforded by item-variant use. Two simulations were performed in an attempt to characterize the value of intrafamily similarities in an easily interpretable context. The simulations approximate the results of administering item variants operationally within a linear testing format without pretesting using only information available from earlier calibrations of other items, at least one of which (in this case the parent) is from the same variant family.

Each simulated examinee was administered a 26-item analogy test and a 36-item antonym test. Each test was created by randomly selecting a single child item from each variant family. For those families that included more than one child item (12 of the analogy families and 13 of the antonym families), different simulated examinees received different items. For those families that included only one child item (14 of the analogy families and 23 of the antonym families), each simulated examinee received the same item. (But recall that the parent and child were not designated systematically; the parent was more difficult in some cases, the child in others — and by amounts that varied from one variant family to the next).

A total of 100 examinees were simulated at each ability value \( \theta \) from -3 to 3 in intervals of .2. This resulted in a total of 3,100 simulated examinees in each simulation. Examinee responses were then simulated according to the calibrated IRT item parameters for the actual items that were administered. All of the examinees were assigned three Bayes-modal ability estimates based on their simulated responses and a standard-normal prior

1. The first estimate was based on the true item parameters — namely, the parameters used to generate the responses.

2. A second estimate was based on the item parameters predicted from the MMR — that is, the mean item parameter vector for all the items in the variant family.

3. The final estimate used expected-response-function (ERF) versions of the predicted item parameters to compensate for the uncertainty in parameter estimates. (The
following paragraphs describe ERFs and discuss how they were generated in these analyses.)

Unlike item-parameter estimates based on calibrations over hundreds or thousands of examinee responses, the parameters predicted by the MMR (the second estimate, above) contain considerable uncertainty. In addition to the usual uncertainty incurred from having only a finite number of calibration responses, additional uncertainty arises from the variation in true parameters among the items in the same variant family. Lewis (1985) introduced expected response functions to handle uncertainty from both of these sources. He noted that the expected probability of a correct response, \( F^*(\theta) \), given a known ability and imperfectly known item parameters, could be expressed as follows:

\[
F^*(\theta) = E_\beta[F(\theta)] = \int P(X = 1 | \theta, \beta) p(\beta) \, \beta.
\]

where \( E_\beta[F(\theta)] \) is the point-wise expectation (conditional on \( \theta \)) of the item response function \( F(\theta) \) over possible values of the item-parameter vector \( (\beta) \), \( F(\theta) = P(X = 1 | \theta, \beta) \) (the IRF), and \( p(\beta) \) is the density function expressing current knowledge about the item-parameter vector. This is the standard Bayesian treatment of uncertainty — that is, integrating over the distribution of the variable(s) at issue (in this case, the item parameters).

The ERF approach differs from the usual approach of treating point estimates of item parameters as known true values. The approaches can be contrasted in terms of the role that item-parameter posterior distributions from the item calibrations play. The usual approach uses the posterior distributions (or likelihood function) to calculate parameter estimates — the expectation of the posterior distribution (or the maximizing value) for each of the item parameters — to produce an IRF used in subsequent inference as if it gave true values of item response probabilities. The ERF approach takes the expectation of the IRF itself (as opposed to expectations of the parameters of the IRF) by integrating over the combined multidimensional posterior distribution for all the item parameters. The ERF approach generally produces a flatter (lower-discrimination) response function than the traditional approach, effectively reducing the certainty placed in the effective IRF.

For the 3PL, Equation 1 requires three-dimensional integration. Mislevy, Wingersky, and Sheehan (1994) proposed a computational approximation similar to the repeated-sampling
approach that Thissen and Wainer (1990) used to characterize the error in maximum-likelihood item parameter estimates. Mislevy, Wingersky, and Sheehan begin with a multivariate-normal approximation of the posterior distributions of the parameters of a given item, from which a large number of \((a, b, c)\) samples are drawn to form an empirical approximation. An IRF is plotted that corresponds to each sampled triple of item parameters, and a point-wise expectation of the resulting IRFs is then calculated. This estimated curve provides the expected value of a 0/1 response as a function of \(\theta\). Mislevy, Wingersky, and Sheehan found that a 3PL curve generally provided a good approximation to these curves, so the mathematical form of the 3PL IRF can be maintained even while accommodating uncertainty in the item parameters. This was the approach taken in the present study for the third ability estimate for a simulated examinee. In particular, for calculating an ERF for each variant in the simulated tests, the \((a, b, c)\) predicted from the MMR for a given item’s variant family was used as the mean, and the common residual covariance was used as the covariance.

Figure 2 shows the mean estimated \(\theta\) against true \(\theta\) for the three ability estimates for analogy and antonym variants. The impact of the standard-normal prior is indicated by the usual inward estimation bias: negative \(\theta\) s tend to be overestimated and positive \(\theta\) s tend to be underestimated. The impact is larger for the shorter analogy test than for the longer antonym test. The \(\theta\) s based on the true item parameters are consistently more accurate than the others, as would be expected, but the differences among the three ability estimates are relatively small over much of the ability range. The ERF ability estimates are always more similar to the true-parameter estimates than are the MMR estimates, indicating the value of using the ERFs rather than IRFs based on the optimal predictions of item parameters.

Figure 3 shows the root mean squared error (RMSE) against true \(\theta\) for the three ability estimates for analogy and antonym variants. The large RMSE values for low- and high-ability examinees indicate the relatively poor measurement obtained outside the middle ability range. This underscores the fact that the test items were not selected for optimal measurement; indeed, these items were never intended to serve as an intact test. Attention should be focused not on the overall quality of measurement but on the differences in quality among the three scoring rules. That the three ability estimates produced nearly indistinguishable RMSEs for middle-ability examinees is comforting, since it is only within this range that relatively good measurement is being accomplished but that most of the action is meant to take place in CAT.
Figure 2. Mean estimated theta using true, multivariate multiple regression analysis and expected-response-function item parameters for analogy and antonym variant simulations.

Figure 3. Root mean squared error using true, multivariate multiple regression analysis and expected-response-function item parameters for analogy and antonym variant simulations.

By how much does having only ERFs, rather than true IRFs, degrade the estimation of examinee abilities? One way to measure the impact is by examining the average per-item contribution to information about $\theta$. Since a standard-normal prior was used for estimating $\theta$, the prior variance for any simulee is 1. Its reciprocal is called the prior precision; in this case, its value is also 1. The reciprocal of the posterior variance is the posterior precision, and is generally greater than the prior precision because of the information contained in the item response vector. Dividing this difference by test length gives an approximation of the average per-item
contribution to information about $\theta$, and calculating the ratio between the IRF and the ERF estimation conditions gives an approximation of the relative value of information between the two conditions. Averaging over a standard normal distribution for $\theta$ yields an average per-item efficiency ratio:

$$\text{Relative Efficiency} = \left( \sum_{q=1}^{31} \phi(\Theta_q) \right)^{-1} \sum_{q=1}^{31} \phi(\Theta_q) \left[ \frac{1}{100} \sum_{i=1}^{100} Var^{-1}(\theta|x_{qi};\text{ERFs}) - 1 \right]$$

$$\left[ \frac{1}{100} \sum_{i=1}^{100} Var^{-1}(\theta|x_{qi};\text{true }\beta\theta) - 1 \right]$$

(2)

where $\Theta_q$ is the $q$th grid point, $\phi(\Theta_q)$ is the standard normal ordinate at that point, $x_{qi}$ is the response vector of the $i$th simulee at the $q$th grid point, $Var^{-1}(\theta|x_{qi};\text{true }\beta\theta)$ is the reciprocal of the posterior variance for $\theta$ given $x_{qi}$ as calculated with true item parameters, and $Var^{-1}(\theta|x_{qi};\text{ERFs})$ is the corresponding value as calculated with expected response functions.

The resulting values for the analogy and antonym simulation data in this study were .93 and .91, respectively. Compared with responses to items with known parameters, then, sample size increases of about 10% would be needed to get the same amount of information about $\theta$ if only ERFs based on variant-family membership alone were available and responses were collected within a linear testing framework.

**Discussion**

Both the bivariate correlations and multivariate regression analyses indicate that the members of variant families tend to be similar in difficulty. Relationships for discrimination and lower-asymptote parameters appear to be much less reliable. Simulation results demonstrate that the intrafamily relationships lead to a decline of only about 10% in precision for examinee $\theta$'s when using variants without pretesting — that is, when predicting item parameters from the parameters of other family members rather than from standard pretesting. The smaller biases and RMSEs of ability estimates based on ERFs rather than on direct predictions from the MMR underscore the importance of accounting for the remaining uncertainty in subsequent inferences.

The ERF- and MMR-based scores were compared against an ideal standard in these simulations. The scores based on true parameters did not reflect the calibration error that would
normally affect score quality; that is, the true-parameter scores are more accurate than scores based on calibrated item parameters would be. On the other hand, the ERFs were “about right” on the average, having been selected at random from the variant families from which the ERFs were constructed. This property cannot be assured if new variant items are created in a given family without a principled algorithm for generating them. For this reason, the use of ERFs without any pretesting seems unadvisable if the variants result from an informal process of production (like the ones studied here).

It is prudent to gather some pretest response data to supplement the information about a new variant’s parameters that is conveyed by the predictive distribution for parameters of its family members. With perhaps a hundred responses rather than several hundred or a thousand, the examiner can in this way produce tailored ERFs that handle unanticipated shifts in difficulty, and can identify the occasional anomalous variants that should be expunged from operational use. Also, the comparison reported in this study used a simulated fixed-form test as opposed to adaptively selected items. Even if errors in ERFs average to zero across the range of difficulty, systematic underestimates or overestimates for hard items or for easy items could introduce biases into examinee estimates in adaptive testing. This possibility would need to be investigated before operational use.

The present study demonstrates that useful information can be gleaned from intrafamily similarities in item functioning, even under the unfavorable conditions of knowing little about variants’ genesis other than that they are variations on common themes. As such, it is an important step toward evaluating more detailed aspects of item-variant production. Future work should proceed along three lines. First, the preliminary work described here would be much supported by additional operational samples. This would allow for some evaluation of the shrinkage in amount of variability explained. Analyses of using ERFs for variants in adaptive testing would need to be included if operational use in this context is foreseen.

Second, future work should include more detailed analyses and systematic manipulations of variations among items. Enright, Morley, and Sheehan’s (2002) work with quantitative-reasoning variants serves as a good example. Through systematic manipulation of item features, the impact of each feature on item functioning can be evaluated within an experimental context. For example, relevant features for analogies may be obtained from a long and distinguished history of cognitive and psychometric analyses of performance in these kinds of tasks (e.g.,
Finally, these more detailed analyses should support work on targeted item production. The item writers who produced the variants used in this study were not attempting to produce items with particular operating characteristics, but detailed knowledge of the relationship between item features and item functioning should support such endeavors in the future. Targeted item production could be extremely useful in meeting the needs for constraint-satisfying items that span ranges of difficulty.
References


