

The Effect of Option Homogeneity in Multiple-Choice Items

Applied Psychological Measurement
2019, Vol. 43(2) 113–124
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618770803
journals.sagepub.com/home/apm



Gregory M. Applegate¹ , Karen A. Sutherland²,
Kirk A. Becker³ and Xiao Luo⁴

Abstract

Previous research has found that option homogeneity in multiple-choice items affects item difficulty when items with homogeneous options are compared to the same items with heterogeneous options. This study conducted an empirical test of the effect of option homogeneity in multiple-choice items on a professional licensure examination to determine the predictability and magnitude of the change. Similarity of options to the key was determined by using subject matter experts and a natural language processing algorithm. Contrary to current research, data analysis revealed no consistent effect on item difficulty, discrimination, fit to the measurement model, or response time associated with the absence or presence of option homogeneity. While the results are negative, they call into question established guidelines in item development. A hypothesis is proposed to explain why this effect is found in some studies but not others.

Keywords

item writing, item difficulty, item analysis

One of the continuing challenges of developing professional educational and psychological assessments is to develop test items that fall within a specified range of difficulty. The ability to modulate the relative difficulty of items by following a blueprint for item construction would enable item developers to target gaps within item banks, increase item survival rates, and enhance the development of algorithms for automatic item generation. One current recommendation to modulate the difficulty of multiple-choice items is to focus on developing options that are more similar or less similar to the key. Items where all options are similar to the key (homogeneous) theoretically should be more difficult and discriminate better than items where the options are less similar to the key (heterogeneous). The purpose of this study is to quantify the changes in item parameters associated with a change in option homogeneity. It expands on

¹National Registry of Emergency Medical Technicians, Columbus, OH, USA

²Oncology Nursing Certification Corporation, Pittsburgh, PA, USA

³Pearson VUE, Chicago, IL, USA

⁴National Council of State Boards of Nursing, Chicago, IL, USA

Corresponding Author:

Gregory M. Applegate, National Registry of Emergency Medical Technicians, 6610 Busch Blvd., Columbus, OH 43229-1797, USA.

Email: gapplegate@nremt.org

previous work by attempting to quantify the difference in item difficulty or discrimination that can be correlated with increasing or decreasing the similarity of options.

Creating homogeneous options has become an accepted guideline for writing multiple-choice items. Haladyna, Downing, and Rodriguez (2002) conducted a survey of 27 textbooks on educational testing and found widespread support for the creation of homogeneous options. Eighteen of the books reviewed supported the use of homogeneous options, whereas the remaining nine neither supported nor refuted the practice. Haladyna et al. (2002) recommended, as a guideline for best practices in item writing, that options that are similar to the key should improve discrimination of the item. However, they qualified the recommendation because not enough empirical evidence existed to definitively support or refute the rule (Haladyna et al., 2002).

Data from at least four research studies support using homogeneous options (Ascalon, Meyers, Davis, & Smits, 2007; Green, 1984; Guttman & Schlesinger, 1966; Smith & Smith, 1988). The original idea of creating homogeneous options comes from Guttman and Schlesinger (1967). They theorized that creating options that are similar to the item key (homogeneous options) in a multiple-choice item would result in items that are more difficult and better discriminating than items in which options are less similar to the item key (heterogeneous options). Their theory was based on a logical analysis of how candidates responded to items and the results of a set of ability tests given to 637 students in Grades 7 through 9 (Guttman & Schlesinger, 1966).

Green (1984) also found a correlation between item difficulty and option homogeneity which she referred to as option convergence. A sample of 19 items from undergraduate students from the University of Washington was analyzed and data suggested that greater similarity among options produced items that were more difficult. This result was obtained by creating nine variations of each item with differences in option convergence and language difficulty. Approximately 300 responses were obtained for each variation, and an ANOVA produced a significant result for option convergence ($\alpha = .01$). The effect for language difficulty was not significant.

As part of a study comparing the Angoff and Nedelsky standard-setting methods, Smith and Smith (1988) used factor analysis to develop a model of item difficulty that included a factor labeled as "Choice Similarity." This factor included two components: similarity and plausibility. Similarity describes a measurement of how closely options resemble the correct choice. Similarity was measured by analyzing the responses of 30 graduate students using a 6-point scale. Each incorrect option was rated on the scale, and the mean of the ratings was used as the value for the item (interrater reliability = .86). Plausibility was a measurement of the degree to which an incorrect response was reasonable. This was measured by four subject matter experts using a 4-point scale (interrater reliability = .67). Smith and Smith also conducted a regression analysis that found a statistically significant result for "Choice Similarity." The 64 items used in this study were obtained from a statewide high-school graduation examination.

Ascalon et al. (2007) used a sample of 493 high-school students who were 13 to 18 years old as well as four examination forms to evaluate the value of similar or dissimilar options. They created 16 items for a California driving examination. Each variation of the examination consisted of four variants of the same item: (a) similar options, (b) dissimilar options, (c) open-ended stem, and (d) question in the stem. Each form contained all 16 items (balanced for difficulty). Thirteen graduate students used a 5-point scale to rate each item set as similar or dissimilar (interrater reliability = .62) and reviewed the items to ensure there was only one correct answer per item. Based on ratings from 13 psychology graduate students, the 16 most homogeneous and 16 most heterogeneous items were selected for data collection. Using a classical test theory analysis, they found that items with more homogeneous options were on average .12 more difficult than were similar items with heterogeneous options.

<i>Heterogeneous</i>	<i>Homogeneous</i>
The nurse is assessing a client with asthma.	The nurse is assessing a client with asthma.
Which of the following findings would the nurse be most likely to observe in a client with asthma?	Which of the following findings would the nurse be most likely to observe in a client with asthma?
a. Wheezing (key)	a. Wheezing (key)
b. Hyperthermia	b. Cough
c. Nausea	c. Crackles
d. Constipation	d. Dyspnea

Figure 1. Example of heterogeneous and homogeneous option sets.

At least one group of researchers found a different result. In 1991, Downing, Dawson-Saunders, Case, and Powell found no significant differences in difficulty or discrimination for items used on exams by the National Board of Medical Examiners (as cited in Haladyna et al., 2002).

Within this small number of research studies, most found that option homogeneity in multiple-choice items did affect at least some item parameters when the items with homogeneous options were compared to items with the same stem but heterogeneous options. One study found that items increased in difficulty by .12 (classical test theory). If the level of option homogeneity is correlated with item difficulty, item developers would have a useful technique for adjusting item difficulty. This would enable test developers to create items that are better matched to a passing standard or to the population of examinees in terms of item difficulty. The ability to create items that more closely match a passing standard or that more closely match the ability distribution of examinees could result in a lower standard error of measurement for a given number of items resulting in shorter or more reliable tests. This information might also prove useful for reducing the cost of producing quality test items.

The literature has used the term “homogeneous items” in at least two contexts. In one context, the item options are similar in length and grammatical construction but not necessarily in content (see Haladyna, 2004, p. 116). In the other, options are similar in length, grammatical construction, and content (see Haladyna & Rodriguez, 2013, p. 105). For the purposes of this study, only the option content was manipulated (see the Figure 1). The length of the distractors and the grammatical structure were kept as consistent as possible using a style guide and visual review of the options.

The heterogeneous example differs from the homogeneous example in that all the options contain symptoms that are unrelated to asthma or respiratory disease except for the key. Hyperthermia (an elevated body temperature), nausea (a feeling of sickness or an inclination to vomit), and constipation (difficulty in emptying the bowels) are all symptoms of diseases or conditions that occur in other systems of the body. A candidate who understands that asthma is a respiratory condition should be able to discern the correct answer by the process of

elimination (the only respiratory symptom) in the heterogeneous example. All of the options in the homogeneous example are symptoms that occur in the respiratory system, so the examinee would need to discern which respiratory symptom would most likely occur in an individual with asthma specifically.

Method

This experiment was conducted using items from a large national professional licensure examination. Item statistics were established by seeding the experimental items with regular items administered during a computerized adaptive test. The experimental items were not administered adaptively and were not distinguishable by the examinees. The experimental item responses were documented but were not counted as part of the test score. At least 533 responses were recorded for each item in the study.

Participants

The test is administered to approximately 200,000 examinees per year. Experimental items were not identified as such but were randomly assigned to examinees during the examination. The experimental items were administered as part of a larger pool of items given to examinees as part of a computerized test. Each examinee received items randomly chosen from the pool of items. As a result, all participants did not see all the experimental items. All responses came from examinees who were educated in U.S. schools of higher education and were testing for professional U.S. licensure. Most examinees were educated at the baccalaureate level. The ethnic self-report information for examinees was African American 9.1%, Asian 10.7%, Caucasian 55.6%, Hispanic 6.1%, Native American 0.5%, Pacific Islander 1.2%, other or not reported 16.8%. The Gender self-report information was Female 83.6%, Male 12.7%, and not reported 2.7%.

Measures

The following information was recorded for each source and variant item.

- Number of responses—a count of the number of examinees who responded to each item.
- Item difficulty in logits—the difficulty of the item measured using a dichotomous Rasch model on an anchored scale.
- Point-measure correlation—the correlation between the estimated ability of the examinees with their responses to each item. See the item discrimination topic in the “Results” section for a discussion of the limits of the point-measure correlation in this context.
- Weighted standardized mean-square fit (infit)—an estimate of how well the item responses fit the expectation of the measurement model. More extreme values (positive or negative) indicate a response pattern that does not meet the assumptions of the model. Infit is a weighted measure that is more sensitive to inlier values (Linacre, 2015). Significant changes to model fit could indicate a change in the measurement value of the item.
- Unweighted standardized mean-square fit (outfit)—an estimate of how well the item responses fit the expectation of the measurement model. More extreme values (positive or negative) indicate a response pattern that does not meet the assumptions of the model. Outfit is an unweighted measure which makes it more sensitive to outlier values than

infit (Linacre, 2015). Significant changes to model fit could indicate a change in the measurement value of the item.

- Response time in seconds—a count of how long an examinee spent in the screen with the item. Measured, in seconds, from the time the screen loaded until the examinee moved to the next screen. This variable was included to determine whether candidates needed more time to select between homogeneous options.

Procedures

A cosine similarity index (CSI) was used to do the initial sort of items into similar or dissimilar option sets. CSI is a text-vector indexing technique which measures the similarity between two vectors of co-occurring text. A distance score is used to compare the similarity of the text (Becker & Kao, 2009; Latifi, Gierl, Wang, Lai, & Wang, 2016). The wording of the key and each option were compared with a root word dictionary. The dictionary provides a reference to collect information about the number of times word roots appear in a section of text. Connecting words such as conjunctions are not used as part of the analysis.

Using the root word dictionary, the words cardiology, cardiac, myocardial, cardiovascular, electrocardiogram, electrocardiographic, and pericardial all count as the same word for purposes of calculating the CSI. The frequency of the root word occurrence produces a vector that can be compared with the vector of another selection of text as expressed in Equation 1.

$$\cos(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|}. \quad (1)$$

CSI values range from 0 (no overlap of words) to 1 (the two sections of text are semantically equal).

An item bank with thousands of operational items was used as the basis for this study. Each item in the bank was analyzed using CSI to do a pairwise comparison of the key to each of the incorrect options. This created a set of three scores for each item. Items with three high scores were considered for inclusion in the homogeneous sample and items with three low scores were considered for inclusion in the heterogeneous sample. The algorithm was used to provide an initial sort of the items but the final determination of the degree of the homogeneity of the option set was made by subject matter experts.

The 100 items with the smallest differences and the 100 items with the largest differences, as determined by the natural language processing algorithm, were reviewed by two subject matter experts who are also trained item developers with at least 10 years' experience each in item development. The subject matter experts selected 50 items from the homogeneous list (content in the key was similar to the content in the other three options) and 50 items from the heterogeneous list (content in the key was dissimilar to the content in the other three options) to use as the experimental sample source items. Only the items in which there was agreement between both subject matter experts were selected.

A variant item was created from each source item. Variant heterogeneous items were created from each homogeneous source item, and a homogeneous variant was created for each heterogeneous source item. In all cases, the stem and key were kept the same and the key remained in the same position relative to the distractors. Distractors were revised following program guidelines to ensure that they were plausible and the item style adhered to established guidelines for best practices in item development. The issue of plausibility is key, as one of the findings of Ascalon et al. (2007) was that plausibility was strongly correlated with option similarity.

This resulted in two samples. One where items with homogeneous option sets had variants created that had heterogeneous option sets and a second sample where items with heterogeneous option sets had variants created with homogeneous option sets. The two-sample design was used as a validity check on significant results. If a significant change was found in one sample, the second sample should have the same effect in the opposite direction. Finding a significant result in only one sample suggests either a statistical artifact or a weak effect.

The variant items were administered on an examination to gather item responses over a period of several months. The items were not identified to the examinees in any way, and the items were delivered randomly, intermixed with regular test items. Items were selected at random for each examinee from the pool of available items by a computer algorithm.

Analysis

A dichotomous Rasch model was used to estimate the item parameters and statistics using an anchored scale to ensure that estimates would be comparable. The Rasch model was used for the analysis because the existing anchored scale and the original item parameters were estimated using the Rasch model. The scale was anchored by using the known item statistics of the operational items that were administered alongside the experimental items. Five item statistics were estimated for each item: item difficulty (in logits), point-measure correlation, standardized weighted mean-square, standardized unweighted mean-square, and response time in seconds.

A one-tailed, paired *t*-test analysis was used to determine whether the mean values for each statistic were significantly different. A one-tailed test was used because the research question is focused on changing the item parameters in a specific direction. The overall alpha used was .05 which was adjusted to .01 to account for family-wise error (multiple comparisons) for five comparisons with the same sample (Dunn, 1961).

The data were also examined by plotting the values for each statistic using the source value as one axis and the variant value as the second axis. A preponderance of points above or below a line where the two values would be equal would provide evidence of a systematic effect.

Part of the analysis is built into the experimental design. The two-sample design was chosen to assist in determining whether results were spurious or systematic. A significant result for one sample should be mirrored (in the opposite direction) in the other sample.

Results

Samples

Fifty items were initially included in each sample; however, due to issues in administration, there were insufficient responses for analysis for some of the items. Results for 47 items from the similar to dissimilar sample and for 49 items from the dissimilar to similar sample are reported. A minimum of 533 responses were recorded for each item calibration, with a mean of 580 responses per item.

Item Difficulty

For the similar to dissimilar sample, the mean difficulty for the source items was $-.31$ and the mean difficulty for the variant items was $-.54$, an absolute difference of $.23$ logits. The *p* value for the one-tailed paired *t* test was $.07$, which is greater than the critical value of $.01$. Figure 2 shows a scatterplot with the source item difficulty and the variant item difficulty for each item pair. Changes were not consistent in direction or magnitude.

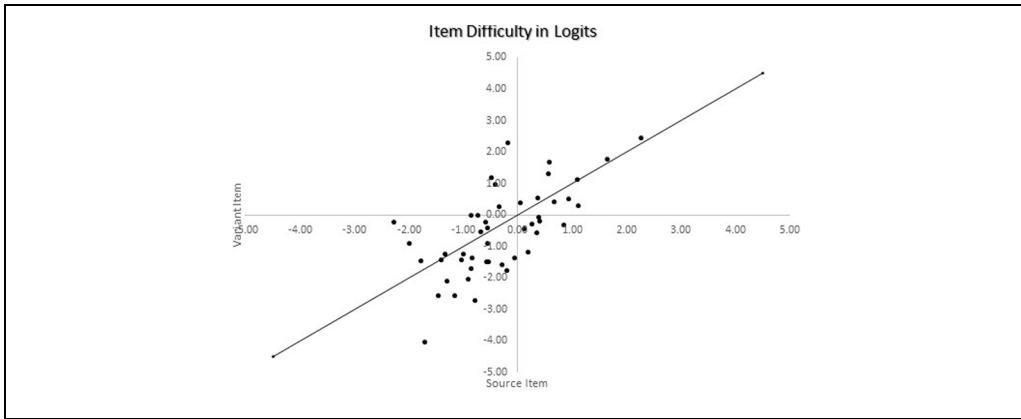


Figure 2. Item difficulty scatterplot for similar to dissimilar sample.

Of the 47 items in the sample, 21 produced difficulty estimates .5 or more logits easier than the source item, 10 produced item difficulty estimates .5 or more logits harder than the source item, and 16 produced item difficulty estimates that were within .5 logits of the source item. Dots above the reference line indicate values that are consistent with the hypothesis that increased option homogeneity correlates with increased item difficulty. Dots below the reference line represent values that refute the hypothesis.

For the dissimilar to similar sample, the mean difficulty for the source items was $-.45$ and the mean difficulty for the variant items was $-.49$, an absolute difference of .04 logits. The p value for the one-tailed paired t test was .42, which is greater than the critical value of .01. Figure 3 shows a scatterplot with the source item difficulty and the variant item difficulty for each item pair. Changes were not consistent in direction or magnitude.

Of the 49 items in the sample, 13 produced difficulty estimates .5 or more logits easier than the source item, 17 produced item difficulty estimates .5 or more logits harder than the source item, and 19 produced item difficulty estimates that were within .5 logits of the source item. Dots below the reference line indicate values that are consistent with the hypothesis that increased option homogeneity correlates with increased item difficulty. Dots above the reference line represent values that refute the hypothesis.

Item Discrimination

For the similar to dissimilar sample, the mean point-measure value for the source items was .12 and the mean point-measure value for the variant items was .10, an absolute difference of .02. The p value for the one-tailed paired t test was .06, which is greater than the critical value of .01. Figure 4 shows a scatterplot with the source item point-measure correlation and the variant item point-measure correlation for each item pair. Changes were not consistent in direction or magnitude.

It may be noted that the values of the point-measure correlations reported here seem small in comparison to other examinations. Discrimination values for the items calibrated with professional certification/licensure populations tend to be lower particularly when compared to K-12 populations due to the broad range of academic preparation (different programs emphasize different parts of the curriculum) and range restriction of ability due to self-selection of the population. Certification/licensure examinees have typically just completed an instructional program

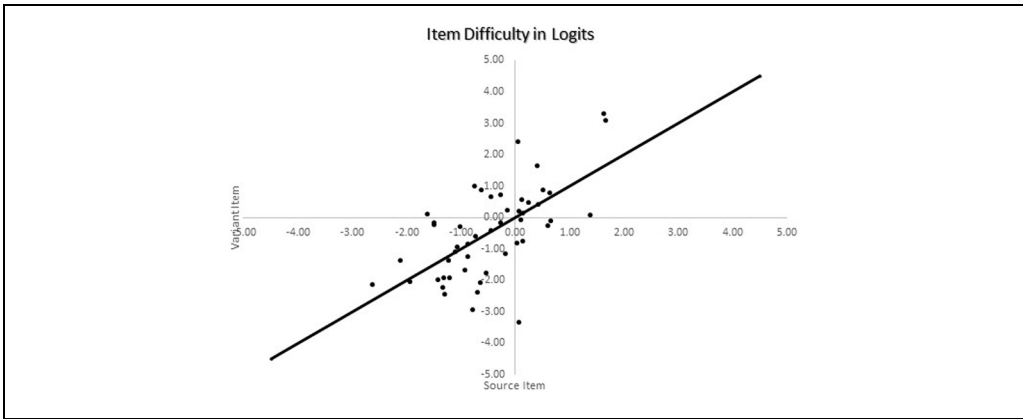


Figure 3. Item difficulty scatterplot for dissimilar to similar sample.

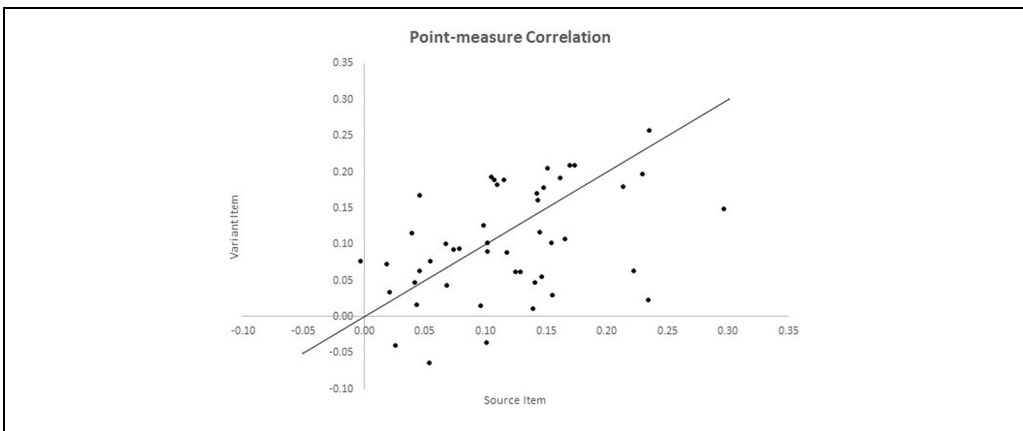


Figure 4. Item discrimination scatterplot for similar to dissimilar sample.

intended to teach them the knowledge they need to become certified or licensed. The instructional focus on the test content and the wide variety of instructional programs make it difficult to achieve high discrimination values for these types of examinations. Item discrimination values in this study are consistent with the item discrimination values of operational items for the same examination.

Of the 47 items in the sample, one produced point-measure estimates which were .1 more discriminating than the source item, seven produced point-measure estimates that were .1 less discriminating, and 39 produced point-measure estimates that were within .1 of the source item. Dots above the reference line indicate values that are consistent with the hypothesis that increased option homogeneity correlates with increased item discrimination. Dots below the reference line represent values that refute the hypothesis.

For the dissimilar to similar sample, the mean point-measure value for the source items was .14 and the mean point-measure value for the variant items was .09, an absolute difference of .05. The p value for the one-tailed paired t test was .0001, which is less than the critical value of .01. Figure 5 shows a scatterplot with the source item point-measure correlation and the

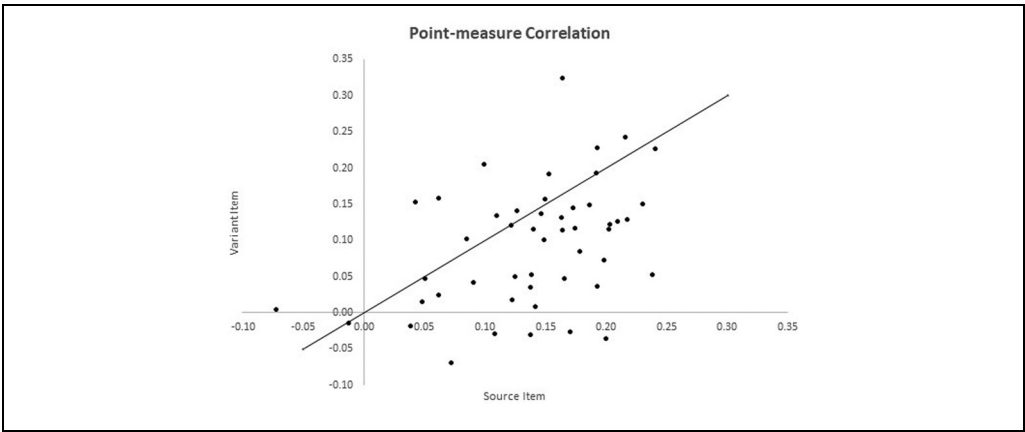


Figure 5. Item discrimination scatterplot for dissimilar to similar sample.

variant item point-measure correlation for each item pair. Changes were somewhat consistent in direction but not in magnitude.

Of the 49 items in the sample, three produced point-measure estimates which were .1 more discriminating than the source item, 12 produced point-measure estimates that were .1 less discriminating, and 34 produced point-measure estimates that were within .1 of the source item. Dots below the reference line indicate values that are consistent with the hypothesis that increased option homogeneity correlates with increased item discrimination. Dots above the reference line represent values that refute the hypothesis.

Summary Table

A similar analysis was conducted on the weighted and unweighted standardized mean-square values and the response times for each item pair. Two statistical tests have significant results, the dissimilar to similar Point-measure comparison and the similar to dissimilar Response Time comparison. A summary of the differences in item statistics is presented in Table 1.

Analysis of Results

This analysis produced two statistically significant results. The mean difference in the point-measure correlation value of .05 for the dissimilar to similar sample runs counter to the theory that more homogeneous item sets produce items that are more discriminating. Not only does this run counter to the theoretical concept but also there is no corresponding change in the similar to dissimilar sample. In the similar to dissimilar sample, the point-measure correlation had a mean difference of .02. The change was not statistically significant.

In addition to the inconsistent effect, it was noted that eight variants in the dissimilar to similar sample had negative point-measure correlation values, whereas the similar to dissimilar sample had four variants with negative point-measure values. All original items had positive point-measure correlation values. A review of the items did not show a consistent reason why. Given the range restriction issues with this population and the small number of items affected in absolute terms, there is not enough evidence to support a theory about the reason for the difference.

The other statistically significant finding was a reduction in response time per item for the similar to dissimilar sample. The response time was 3.6 s less for the dissimilar items but again,

Table 1. Summary of Mean Differences Between Samples (Source Value Minus Variant Value).

Similar to dissimilar					
	Item difficulty (logits)	Point-measure	Weighted standardized mean-square	Unweighted standardized mean-square	Response time
<i>M</i>	0.23	.02	0.02	−0.03	3.59*
Median	0.37	.00	0.17	0.06	5.45
Minimum	−2.45	−.12	−4.44	−4.51	−21.61
Maximum	2.34	.21	3.45	3.62	21.04
<i>SD</i>	1.02	.08	1.51	1.52	8.82
Dissimilar to similar					
	Item difficulty (logits)	Point-measure	Weighted standardized mean-square	Unweighted standardized mean-square	Response time
<i>M</i>	0.03	.05*	−0.37	−0.50	2.54
Median	0.00	.05	−0.13	−0.23	2.53
Minimum	−2.34	−.16	−3.71	−4.04	−16.16
Maximum	3.43	.24	3.58	3.57	26.68
<i>SD</i>	1.12	.08	1.60	1.60	9.21

*A statistically significant result using $\alpha = .01$.

this finding was not supported by the dissimilar to similar sample where the similar items has a shorter response time by 2.5 s but this result was not statistically significant. No consistent effect was found for any of the five item parameters studied.

Discussion

The conclusions from this study do not directly address the research question given here because a key assumption, that option homogeneity consistently affects item difficulty, was not true for this dataset. The lack of an effect related to the homogeneity of options was unexpected because four previous research studies had all detected a directional effect, and Guttman and Schlesinger's argument is logically appealing. Notably, the one dissenting research study sampled an adult, college-educated population like the sample used for this study.

Considering the previous studies, the results of this research suggest at least two possible theories. First, the effect may be age related. Previous studies that used undergraduate or younger examinees all found an effect. The Ascalon et al. (2007) and Smith and Smith (1988) studies both used high-school student responses for item calibrations. The Guttman and Schlesinger (1966) study used responses from students in Grades 7 through 9. Green's (1984) respondents were undergraduate students, but the specific academic level of the students was not indicated.

The one study that did not find an effect (Downing et al., 1991) was conducted using items from an examination developed by the National Board of Medical Examiners. The sample responses came from medical students who were testing for professional licensure.

Their results along with those from this study suggest that option homogeneity may have a greater effect on younger examinees and that the effect appears to diminish or disappear in the adult population (or by education level). This change could be attributed to the maturity level of the test taker. An older, more mature individual may recognize the significance of high-stakes testing for licensure which in turn could affect preparation for and focus during the test. The

change could also be attributed to greater experience and understanding of testing (testwiseness) which would be correlated with the age of the test takers. Maturity and testwiseness were not included in this study, so there is no direct evidence to support this theory. Additional research is needed to support this theory.

The lack of an observed effect might also be due to the self-selection of the samples involved. The studies involving public school students are probably reflective of the general population, but the samples involving college-prepared test takers reflect a smaller proportion of the general population and one that has self-selected for academic work.

An alternative theory is that the homogeneity effect found in previous studies may be related to the plausibility of item options. Ascalon et al. (2007) found a strong correlation between option homogeneity and subject matter expert evaluations of the plausibility of options. The items used in this study were independently reviewed by multiple committees of subject matter experts to ensure that all options were plausible. Logically, all other things being equal, items with plausible options should be at least slightly more difficult than items with at least one implausible option.

The results of this study found no consistent correlation between item difficulty and option homogeneity. Given the consistency of this result with previous research using a professional population, changing options to be homogeneous is not a useful method for altering item difficulty for professional certification/licensure examinations at or above the baccalaureate level.

Limitations and Future Research

Several factors limit the generalizability of these findings. The experimental design for this study used existing item calibrations from source items that had been collected at an earlier point in time. In some cases, there may have been a confounding effect from item drift; however, item parameter estimates are regularly checked for drift and item fit, and item parameter estimates have been found to be stable over time for this program.

Although this study was conducted with a reasonably large number of matched items, the sample population was limited to adult, professionally trained college-educated examinees. In addition, the content of all sample items was related to a single construct, similar to that of the Downing et al. (1991) study, and the general content of the items is well known to the examinee population. In both studies, all examinees had prepared to take high-stakes licensure tests, and the nature of a high-stakes test, in and of itself, may have influenced focus during testing and the observed lack of variance in option homogeneity.

Developing an experiment to separate variance in item difficulty due to option homogeneity and option plausibility is challenging. One particular challenge is the definition of plausibility. For future studies, it would be useful to include a step in the process where the plausibility of options are rated independently by subject matter experts or by asking each test taker to rate the plausibility of each option on a scale. A research design that compares item pairs across age groups would be useful for exploring the age-related effect theory.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Gregory M. Applegate  <https://orcid.org/0000-0001-9632-5472>

References

- Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education, 20*, 153-170.
- Becker, K. A., & Kao, S. (2009, April). *Finding stolen items and improving item banks*. Paper presented at the American Educational Research Association Annual Meeting, San Diego, CA.
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52-64.
- Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement, 44*, 551-561.
- Guttman, L., & Schlesinger, I. M. (1966). *Development of diagnostic analytical and mechanical ability tests through facet design and analysis* (Project No. OE-IS-1-64, Office of Education, U.S. Department of Health, Education and Welfare). Jerusalem: Israel Institute of Applied Social Research. Retrieved from <http://files.eric.ed.gov/fulltext/ED010590.pdf>
- Guttman, L., & Schlesinger, I. M. (1967). Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement, 27*, 569-580.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309-334.
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.
- Latifi, S., Gierl, M., Wang, R., Lai, H., & Wang, A. (2016). Information-based methods for evaluating the semantics of automatically generated test items. *Artificial Intelligence Research, 6*, 69-79.
- Linacre, J. M. (2015). *Winsteps® Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps.com.
- Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25*, 259-274.