The effect of 'None of the Above' on multiple choice questions in a first year classroom

Matthew V. Pachai[a], David DiBattista[b], Joseph A. Kim[a]

[a]*Department of Psychology, Neuroscience, and Behaviour, McMaster University, Hamilton, Ontario, CANADA*
[b]*Department of Psychology, Brock University, St. Catharines, Ontario, CANADA*

Abstract

A variety of 'best practice' guides to multiple choice question writing exist in the literature, each commonly containing a set of specific recommendations. One such recommendation entails the use of 'none of the above' (NOTA), with some authors discouraging and others advocating its use on multiple choice tests. Empirical research on the use of NOTA has produced mixed results regarding its effects on a question's difficulty and ability to discriminate between high-performing and low-performing students. Many of these empirical studies are conducted in the laboratory, or in relatively small classrooms. In the current study, we assess the effect of NOTA on question difficulty and discrimination in a large Introductory Psychology classroom. We find that NOTA increases question difficulty, but only when it is used as the correct response to the question. Moreover, we find no effect of NOTA on a question's discrimination ability. These findings serve to clarify the effect of NOTA on multiple choice questions in true classroom environments.

Key words: multiple choice, none of the above, nota, assessment, evaluation

Introduction

Multiple choice tests have become a common form of assessment in undergraduate education, and almost every undergraduate student will complete a multiple choice test at some time during their degree. A typical multiple choice question consists of a stem, where the question is posed to the student, and a series of options. Within the options is the key, which is the correct response to the question, presented with a set of typically two to four distractors. The student receives a mark if they select the key and not a distractor. With a typical multiple choice question containing one key and three distractors, students can guess the correct response 25% of the time even if they have no knowledge of the content. For this reason, a variety of methods have been developed to correct for guessing (Diamond and Evans, 1973; Frary, 1988; Budescu and Bar-Hillel, 1993). Conversely, others argue that guessing should not be a significant concern for test writers because students with a moderate level of engagement in the course material will rarely engage in truly random guessing, instead opting to eliminate distractors and select amongst the remaining responses (Ebel, 1968). In this way, multiple choice testing does, to some degree, test every students' knowledge of the course material.

Given the ubiquity of multiple choice tests, it should not be surprising that a number of "best practice" guides for question writing have been produced (for a review see Haladyna et al., 2002). These guides generally consist of a large number of recommendations, such as posing the stem as a question or avoiding clues to the correct re-

Email addresses: pachaim@mcmaster.ca (Matthew V. Pachai), david.dibattista@brocku.ca (David DiBattista), kimjoe@mcmaster.ca (Joseph A. Kim)

sponse. However, questions with significant flaws remain common on most multiple choice assessments (Jozefowicz et al., 2002; Downing, 2002, 2005). For example, in a sample of examinations distributed to first and second year medical students, (Downing, 2005) found flaws in 46% of the total questions, with the number of flaws per test ranging from 36% to 65%. These flaws were derived from the recommendations of Haladyna et al. (2002) and include unfocused stems, stems worded in the negative, and the use of 'all of the above' among others. Flawed questions are a significant concern, as they have been shown to negatively impact student learning and to disproportionately hinder the performance of the most knowledgeable students over others (Downing, 2005; Tarrant and Ware, 2008).

Another concern regarding guides for multiple choice question writing is that the suggestions therein are infrequently based on empirical research (Frey et al., 2005). Although some progress has been made in recent years, there still remains a clear need for empirical research on common practices in multiple choice test writing. This is the goal of the present study, in the context of one popular multiple choice option - none of the above (NOTA).

Multiple choice test writers are decidedly split in their opinions on the use of NOTA. In their recent review of test writing guidelines, Haladyna et al. (2002) note that 48% of authors believe that NOTA should never be used, while 44% believe that NOTA has its place in multiple choice tests if implemented thoughtfully. Authors who advocate the use of NOTA cite its ability to increase question difficulty as a favourable outcome (Frary1991). One of the most common arguments against the use of NOTA is that it can reward students with no knowledge of the course content. Gross (1994) points out that when NOTA is the key, a student who does not know the true response can select NOTA and receive equal credit to a student who did know the true response. He goes on to argue that any question format that rewards students with incorrect information is inherently flawed and should never be used.

Empirical research has also produced mixed results on the effect of NOTA. Most studies have found that NOTA has the effect of increasing test difficulty (Rimland, 1960; Dudycha and Carpenter, 1973; Forsyth and Spratt, 1980; Crehan and Haladyna, 1991; Oosterhof and Coats, 1984; Tollefson, 1987). However, it is important to distinguish between the presence of NOTA as the key or as a distractor. In many studies demonstrating an effect of NOTA on item difficulty, the difference between NOTA as a key and NOTA as a distractor is not considered, and as a result these studies often demonstrate relatively small effect sizes (Rimland, 1960). In other studies, NOTA has been found to increase difficulty when it is the key, but not when it is a distractor (Tollefson, 1987; Oosterhof and Coats, 1984). Another common measure of multiple choice item efficacy is the discrimination coefficient, which assesses how well a particular item discriminates between high performing and low performing students. In the few studies measuring the effect of NOTA on discrimination, most demonstrate no effect of NOTA, either as the key or as a distractor (Dudycha and Carpenter, 1973; Tollefson, 1987; Crehan and Haladyna, 1991).

Given the relative lack of research on the effect of NOTA, it is clear that more research is required to develop a clear consensus. Although authors such as Gross (1994) suggest that NOTA should never be used, a large contingent of item writing guidelines do advocate for its use when question difficulty is too low (Frary, 1991). Moreover, relatively few studies on the effect of NOTA are conducted in actual classrooms, with researchers instead opting to conduct laboratory experiments due to the added control gained in that setting. However, compared to classroom experiments, laboratory studies are in some ways disadvantaged. Specifically, in the laboratory, the test content must be general, so as to be appropriate for the random sample of participant. If the content is too specialized, it must be learned in the laboratory itself either earlier in the experimental ses-

sion or in an earlier session some time before. While these conditions are somewhat analogous to the tests distributed to students in a classroom, there are a number of key differences between the two settings. During a classroom test, a student's performance on the test is critical to their final grade in the course and subsequently their academic goals at large. Therefore, students in a classroom may be more motivated to select the correct response than participants in a laboratory. Moreover, students will spend more time studying the course content, and will have been immersed in the material throughout the term. In this way, only by conducting classroom studies can researchers hope to make conclusions regarding the use of NOTA in true educational settings.

In the present study, we used of McMaster's large first year Psychology course to assess the effect of NOTA on item difficulty and discrimination in the classroom. By utilizing this sample of students, we are able to take advantage of the more ecologically valid classroom environment with a large sample size. In the present study, we manipulated questions on the midterm examination distributed to Introductory Psychology students each semester. Due to our large sample of students, we can be confident that our estimates of difficulty and discrimination are more stable and accurate than those obtained with small classes that can vary more from year to year. This allowed for a question-based approach to the design of the study, where a specific question was varied across versions of the test such that it appeared without NOTA (control), with NOTA as the key, or with NOTA as a distractor for different groups of students. By using a carefully controlled design in a true classroom populated with a large number of students, we aim to clarify the effect of NOTA on item difficulty and discrimination in a real academic setting.

Methods

Psychology 1X03/1XX3

Psychology 1X03 and 1XX3 together comprise the first year psychology program at McMaster University. Psychology 1X03 is offered in the first term and acts as a prerequisite for Psychology 1XX3, which is offered in the second term. Each year, approximately 3100 students enroll in Psychology 1X03 from all faculties including Science, Social Science, Humanities, Business, and Nursing, making this course the largest on the McMaster University campus. Topics covered in Psychology 1X03 include research methods, learning, memory, cognition, social psychology, and personality theory. Psychology 1XX3 is offered in the second term and serves approximately 1700 students each year. Topics covered in Psychology 1XX3 include development, evolution, neuroscience, and sensory systems. In both Psychology 1X03 and 1XX3, students complete one midterm examination (25% value) and one final examination (40% value) per semester, both in the multiple choice format, along with other evaluation components including class participation and written assignments.

Participants

Participants consisted of students completing their regularly scheduled midterm examination in Psychology 1X03 and 1XX3 in the 2008/2009 and 2009/2010 academic years. Students consented to have their midterm examinations included in this study during the course of their final examination. A consent form was included on the last page of this examination and a bonus mark was offered to students who responded, regardless of the nature of their response. All research protocols were approved by McMaster University's Research Ethics Board. Two midterm examinations in Psychology 1X03 were included in this study, one from the 2008/2009 academic year, and one from the 2009/2010 academic year. One Psychology 1XX3 midterm examination was also included in this

study, administered in the 2008/2009 academic year. The total sample of consenting students was 1955 (63% participation) and 1742 (56% participation) for the two Psychology 1X03 examinations, and 891 (52% participation) for the Psychology 1XX3 examination.

Materials

On each midterm examination, a subset of the total questions was manipulated for inclusion in this study. The 2008/2009 Psychology 1X03 examination included 5 experimental questions embedded in 25 total questions, and the 2009/2010 Psychology 1X03 examination included 10 experimental questions embedded in 35 total questions. The 2008/2009 Psychology 1XX3 examination included 5 experimental questions embedded in 40 total questions. Therefore, we analyzed a total n of 20 experimental questions. On every examination, the experimental questions were randomly distributed throughout the test and were indistinguishable in style from non-experimental questions. Every question consisted of one key and three distractors.

Design

Each of the three midterm examinations consisted of five test versions distributed randomly to the student population. These five versions differed in the numerical placement of the questions to discourage cheating. The non-experimental questions were identical on every test version, while the experimental questions were manipulated across test versions for the purposes of the study. Specifically, the experimental questions differed in the placement of the NOTA option. For a single experimental question, one test version would contain the question in its original format with no NOTA option; one test version would contain the question with NOTA replacing the key; and the remaining three test versions would contain the question with NOTA replacing each of the three distractors respectively. Moreover, after completion of the test, we classified the distractors according how frequently they were selected in the control condition. This yielded three distractor types ranked according to their selection frequency - high frequency (HF), medium frequency (MF), and low frequency (LF). Therefore, in total there were five experimental conditions: Control (no NOTA), NOTA replacing the key, NOTA replacing the HF distractor, NOTA replacing the MF distractor, and NOTA replacing the LF distractor.

To prevent any systematic differences in the difficulty level of the five test versions, the five experimental conditions were also balanced within each version. Specifically, each test version contained one control question, one question with NOTA replacing the key, and three questions with NOTA replacing the three distractors respectively. In this way, the five experimental conditions were balanced across test versions as well as within test versions. Other than the NOTA manipulation, experimental questions were identical to non-experimental questions in style and format, although the non-experimental questions never contained NOTA as an option.

Results

Statistical analyses were done with R (R Development Core Team, 2007). When appropriate, the Huynh-Feldt correction, $\tilde{\epsilon}$, was used to adjust p values of F tests conducted with within-subject variables for violations of sphericity (Maxwell and Delaney, 2004). Effect size was expressed as Cohen's f. All t tests reported were one tailed. Our dependent measures were difficulty and discrimination. Difficulty was defined as the percentage of students who answered a particular question correctly. Discrimination was defined as the point-biserial correlation between the student population's responses for a particular question, either correct or incorrect , and the student population's total score on the examination. Each experimental question was considered to be a subject in a

repeated-measures design.

Figure 1a demonstrates the effect of NOTA placement on question difficulty. These data were submitted to a one-way repeated-measures ANOVA, with NOTA placement as the within-subjects factor. The ANOVA revealed a highly significant main effect of NOTA placement on question difficulty, $F(19, 76) = 16.29$, $\tilde{\epsilon} = 0.67$, $p < 0.0001$, $f = 0.78$. To further analyze these differences, we performed a number of post-hoc linear comparisons between conditions. The first comparison revealed that the NOTA-as-key condition was significantly more difficult than all other conditions combined, $t(19) = 5.23$, $p < 0.00001$. We also found no difference between the control condition and the conditions where NOTA replaces a distractor, collectively, $t(19) = 1.50$, $p > 0.05$. Finally, we found that the effect on difficulty of replacing the high frequency distractor with NOTA was not significantly different than replacing the medium or low frequency distractors with NOTA, $t(19) = 1.26$, $p > 0.1$. Therefore, it appears that replacing the key with NOTA significantly increases question difficulty (lowers percent correct), while replacing any distractor with NOTA has no effect on difficulty relative to when NOTA is not present in the question.

Figure 1b demonstrates the effect of NOTA placement on the point-biserial discrimination scores. The mean point-biserial discrimination score in the control (no NOTA) condition was 0.25, which falls within the range commonly reported in large samples of multiple choice questions (Martínez et al., 2009). These data were submitted to a one-way repeated-measures analysis of variance with NOTA placement as the within-subjects factor. The ANOVA revealed no significant main effect of NOTA placement on discrimination, $F(19, 76) = 1.05$, $\tilde{\epsilon} = 1.03$, $p > 0.3$, $f = 0.045$. Therefore, it appears that NOTA has no effect on discrimination scores regardless of where it is placed in the question.

Discussion

The present study examined the effect of NOTA on multiple choice questions in a large classroom setting. Specifically, we manipulated the placement of NOTA across versions of the test such that, for each of 20 experimental questions, NOTA was either not present, replaced the key, or replaced the high, medium, or low frequency distractors respectively. We found that NOTA significantly increased the difficulty of multiple choice questions only when it was used as the key. We found no effect of NOTA replacing any of the distractors, regardless of their effectiveness. Additionally, we found no significant effect of NOTA on discrimination scores, regardless of its placement.

These findings are consistent with previous studies that have shown an increase in question difficulty with the use of NOTA, and no effect of NOTA on discrimination (Tollefson, 1987; Oosterhof and Coats, 1984; Dudycha and Carpenter, 1973; Crehan and Haladyna, 1991). Moreover, our results affirm the importance of distinguishing between the presence of NOTA as the key or as a distractor, which is not always the case (Rimland, 1960; Oosterhof and Coats, 1984). Interestingly, we observed no significant difference between replacing a high, medium, or low frequency distractor with NOTA. This suggests that students treat the NOTA option similarly regardless of the allure of the other two distractors. However, we do observe a slight trend towards lower difficulty when higher frequency distractors are replaced by NOTA, suggesting that NOTA may not always be as effective as a well written distractor.

Empirical studies on the writing of writing multiple choice questions have become increasingly important in light of a large body of literature suggesting that multiple-choice tests can also be used for the purposes of improving learning (Carrier and Pashler, 1992; Hogan and Kintsch, 1971; Thompson et al., 1978; Roediger and Marsh, 2005; Karpicke and Roediger, 2007). This observation, dubbed the Testing Effect, is derived from exper-

iments demonstrating improvements in later recall when students have previously written a test on some material as opposed to just having studied it. With this observation in hand, instructors can strive to improve student learning by creating challenging multiple choice tests that engage the students and get them thinking about the course content. However, a number of negative consequences of multiple choice testing have also been observed (Roediger and Marsh, 2005; Roediger, H. L., 1996; McDermott, 2006). These negative consequences likely stem from what Remmers and Remmers (1926) dubbed the Negative Suggestion Effect. In this effect, exposure to misinformation can increase the probability of recognizing or recalling that same misinformation on a later test. For example, Jacoby and Hollingshead (1990) found that exposure to misspelled words increased the probability of spelling errors on a later test; while Brown et al. (1999) found that the presentation of incorrect information between an initial test and a later test increased the probability of incorrect responses on the later test. In this way, when students are exposed to the distractors on a multiple choice test, they are being exposed to misinformation that may later be reproduced.

Given the balance between the positive and negative effects of multiple choice testing on learning, small decisions regarding the formatting and style of multiple choice questions can be critical. Generally, the benefits of multiple choice testing are strongest when students are able achieve a high level of performance on the test Roediger and Marsh (2005); Butler et al. (2006). When students select the key, this correct information is reinforced by the test, while students who select distractors are reinforcing incorrect information, as in the negative suggestion effect. Interestingly however, students who receive feedback on their responses, be it immediate or delayed, are less likely to reproduce distractors on a later test (Butler and Roediger, 2008). These students gain information about the correct response can use this information to their benefit on later tests regardless of what they originally selected.

These cognitive considerations apply to every multiple choice question writing decision, including the use of NOTA. Odegard and Koen (2007) have demonstrated that when NOTA is present as the key, but not when it is a distractor, the positive testing effect is negated. They argue that when NOTA is the key, students will either commit themselves to a distractor or they will select NOTA without knowing the true response, both of which lead them to reinforce incorrect information. The significant increases in question difficulty we observe when NOTA is present as the key suggest that a large number of students fall into this trap, however without further research it remains unclear what proportion of students selecting NOTA know the true response to the question and what proportion are consolidating misinformation. Therefore, before making decisions regarding issues such as the use of NOTA, each multiple choice question writer should consider the goals of their assessment tool. On one hand, testing serves the important function of discriminating between students and assigning ranks and grades. To this end, NOTA can serve to increase the difficulty level of a test and challenge the students who have not studied enough. On the other hand, decades of research in cognitive science suggests that tests can serve to enhance student learning. Here, the difficulty increase associated with NOTA may be undesirable, as it can counteract the positive effects of multiple choice questions on learning Odegard and Koen (2007).

It is clear that careful evaluation of every aspect of multiple choice question writing is critical to the effectiveness of the format. The present research has demonstrated a clear and ecologically valid effect of NOTA on difficulty and helped to clarify a field where methodologies and results can vary wildly. Given the relative lack of careful empirical examinations of multiple choice question writing in a classroom context, much future research is needed before we can arrive at a scientifically grounded consensus on the so-called "best practices" for test writing. The development of such a consensus and the careful implementation of the recommendations therein should be a pri-

mary concern for anyone interested in delivering a thoughtful and effective multiple choice evaluation.

References

Brown, A. S., Schilling, H. E. H., Hockensmith, M. L., 1999. The Negative Suggestion Effect: Pondering Incorrect Alternatives May be Hazardous to Your Knowledge. Journal of Educational Psychology 91 (4), 756–764.

Budescu, D., Bar-Hillel, M., 1993. To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring. Journal of Educational Measurement 30 (4), 277–291.

Butler, A. C., Marsh, E. J., Goode, M. K., Roediger, H. L., 2006. When additional multiple-choice lures aid versus hinder later memory. Applied Cognitive Psychology 20 (7), 941–956.

Butler, A. C., Roediger, H. L., 2008. Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. Memory & Cognition 36 (3), 604–616.

Carrier, M., Pashler, H., 1992. The influence of retrieval on retention. Memory & Cognition 20 (6), 633–642.

Crehan, K., Haladyna, T. M., 1991. The Validity of Two Item-Writing Rules. Journal of Experimental Education 59 (2), 183–192.

Diamond, J., Evans, W., 1973. The Correction for Guessing. Review of Educational Research 43 (2), 181–191.

Downing, S. M., 2002. Construct-irrelevant variance and flawed test questions : Do multiple-choice item-writing principles make any difference? Medical Education, 103–104.

Downing, S. M., 2005. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Advances in Health Science Education 10 (2), 133–143.

Dudycha, A. L., Carpenter, J. B., 1973. Effects of item format on item discrimination and difficulty. Journal of Applied Psychology 58 (1), 116–121.

Ebel, R. L., 1968. Blind Guessing on Objective Achievement Tests. Journal of Educational Measurement 5 (4), 321–325.

Forsyth, R. A., Spratt, K. F., 1980. Measuring Problem Solving Ability in Mathematics with Multiple-Choice Items : The Effect of Item Format on Selected Item and Test Characteristics. Journal of Educational Measurement 17 (1), 31–43.

Frary, R. B., 1988. Formula Scoring of Multiple Choice Tests (Correction for Guessing). Educational Measurement: Issues and Practice 7 (2), 33–38.

Frary, R. B., 1991. The none-of-the-above option: An empirical study. Applied Measurement in Education 4, 115–124.

Frey, B., Petersen, S., Edwards, L., Pedrotti, J., Peyton, V., 2005. Item-writing rules: Collective wisdom. Teaching and Teacher Education 21 (4), 357–364.

Gross, L. J., 1994. Logical versus empirical guidelines for writing test items: The case of "none of the above". Health Professions 17 (1), 123–126.

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 (3), 309–334.

Hogan, R., Kintsch, W., 1971. Differential effects of study and test trials on long-term recognition and recall. Journal of Verbal Learning and Verbal Behavior 10 (5), 562–567.

Jacoby, L. L., Hollingshead, A., 1990. Reading student essays may be hazardous to your spelling: Effects of reading incorrectly and correctly spelled words. Canadian Journal of Psychology 44 (3), 345–358.

Jozefowicz, R., Koeppen, B., Case, S., Galbraith, R., Swanson, D., Glew, H., 2002. The quality of in-house medical school examinations. Academic Medicine 77, 156–161.

Karpicke, J., Roediger, H. L., 2007. Repeated retrieval during learning is the key to long-term retention. Journal of Memory and Language 57 (2), 151–162.

Martínez, R. J., Moreno, R., Martín, I., Trigo, M. E., 2009. Evaluation of five guidelines for option development in multiple-choice item-writing. Psicothema 21 (2), 326–30.

Maxwell, S., Delaney, H., 2004. Designing Experiments and Analyzing Data: A Model Comparison Approach, 2nd Edition. Lawrence Erlbaum Associates, Mahwah, New Jersey.

McDermott, K. B., 2006. Paradoxical effects of testing: repeated retrieval attempts enhance the likelihood of later accurate and false recall. Memory & Cognition 34 (2), 261–7.

Odegard, T. N., Koen, J. D., 2007. "None of the above" as a correct and incorrect alternative on a multiple-choice test: implications for the testing effect. Memory 15 (8), 873–85.

Oosterhof, A. C., Coats, P. K., 1984. Comparison of Difficulties and Reliabilities of Quantitative Word Problems in Completion and Multiple-Choice Item Formats. Applied Psychological Measurement 8 (3), 287–294.

R Development Core Team, 2007. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Remmers, H., Remmers, E. M., 1926. The negative suggestion effect of true false examination questions. The Journal of Educational Psychology 17, 52–56.

Rimland, B., 1960. The effects of varying time limits and of using "right answer not given" in experimental forms of the U.S. Navy arithmetic test. Educational and Psychological Measurement 20 (3), 533–539.

Roediger, H. L., Marsh, E. J., Sep. 2005. The positive and

negative consequences of multiple-choice testing. Journal of Experimental Psychology. Learning, Memory, and Cognition 31 (5), 1155–9.

Roediger, H. L., 1996. Misinformation Effects in Recall: Creating False Memories through Repeated Retrieval. Journal of Memory and Language 35 (2), 300–318.

Tarrant, M., Ware, J., Feb. 2008. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Medical education 42 (2), 198–206.

Thompson, C. P., Wenger, S. K., Bartling, C. A., 1978. How recall facilitates subsequent recall: A reappraisal. Journal of Experimental Psychology: Human Learning & Memory 4 (3), 210–221.

Tollefson, N., 1987. A Comparison of the Item Difficulty and Item Discrimination of Multiple-Choice Items Using the "None of the Above" and One Correct Response Options. Educational and Psychological Measurement 47 (2), 377–383.

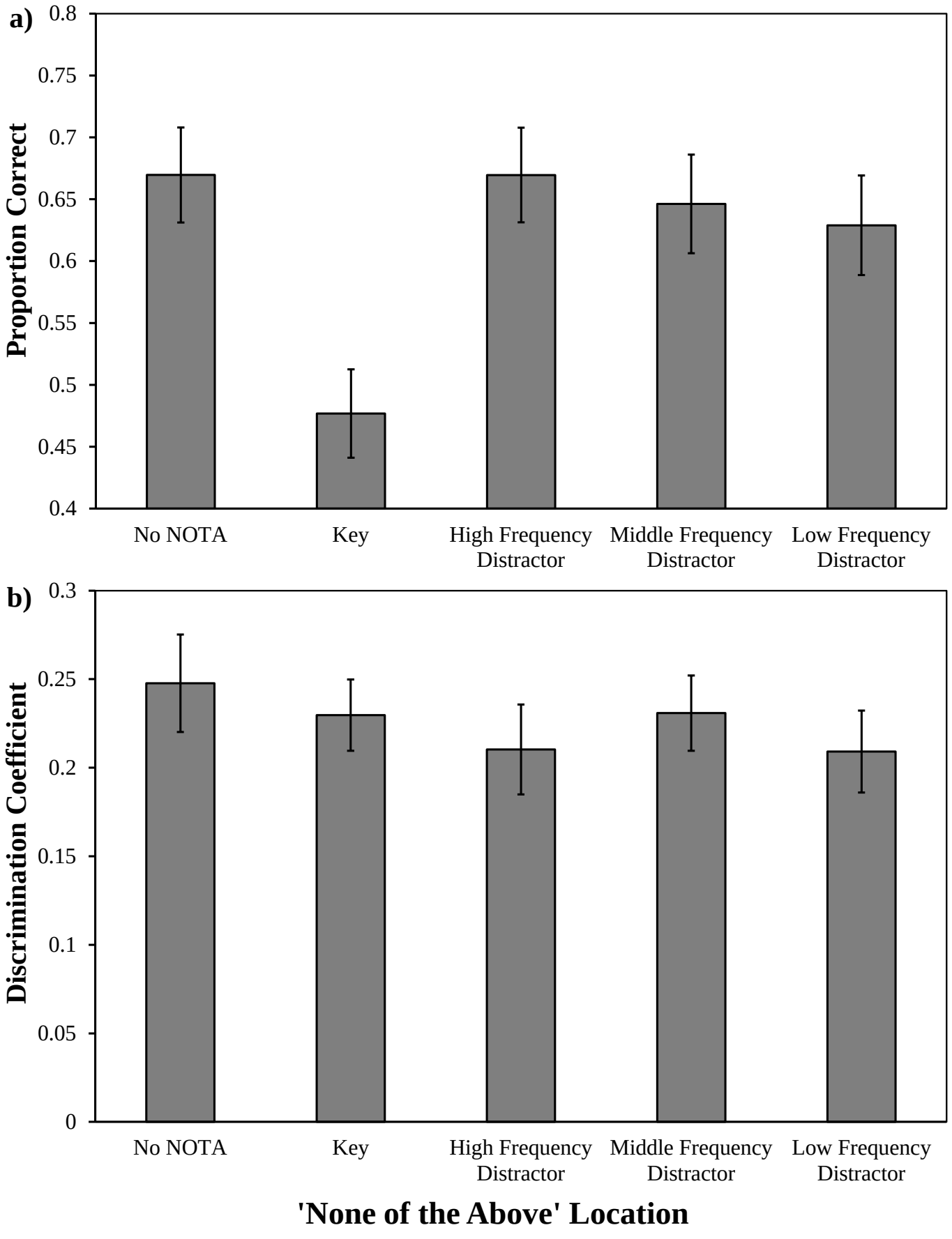
a)
Proportion Correct
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
No NOTA
Key
High Frequency Distractor
Middle Frequency Distractor
Low Frequency Distractor
b)
Discrimination Coefficient
0.3
0.25
0.2
0.15
0.1
0.05
0
No NOTA
Key
High Frequency Distractor
Middle Frequency Distractor
Low Frequency Distractor
'None of the Above' Location

Figure 1: Top: Proportion correct as a function of 'none of the above' placement. Bottom: Point-biserial discrimination scores as a function of 'none of the above' placement