Journal of Educational Measurement
Summer 2003, Vol. 40, No. 2, pp. 109–128

Guess Where: The Position of Correct Answers in Multiple-Choice Test Items as a Psychometric Variable

Yigal Attali
Educational Testing Service
Maya Bar-Hillel
The Hebrew University of Jerusalem

In this article, the authors show that test makers and test takers have a strong and systematic tendency for hiding correct answers—or, respectively, for seeking them—in middle positions. In single, isolated questions, both prefer middle positions to extreme ones in a ratio of up to 3 or 4 to 1. Because test makers routinely, deliberately, and excessively balance the answer key of operational tests, middle bias almost, though not quite, disappears in those keys. Examinees taking real tests also produce answer sequences that are more balanced than their single question tendencies but less balanced than the correct key. In a typical four-choice test, about 55% of erroneous answers are in the two central positions. The authors show that this bias is large enough to have real psychometric consequences, as questions with middle correct answers are easier and less discriminating than questions with extreme correct answers, a fact of which some implications are explored.

"[Ronnie's] grades were okay in junior high because his ear for multiple-choice tests was good—in Lake Wobegon, the correct answer is usually 'c.' "—Garrison Keillor (1997, p. 180).[1]

This article explores the role of within-item answer position in multiple-choice tests. We show that there are strong and systematic position effects in the behavior of both test takers and test makers—even the professionals who produce the SAT—and we explore their psychometric consequences. The article is organized as follows: The first section discusses how people might go about positioning the correct answer to an isolated multiple-choice term. In the second section, we present the empirical evidence on whether writers favor middle positions. The third section shows that test takers favor middle positions as well. Then, we suggest that the bias to the middle is probably really a bias against the edges. We then extend the evidence presented to entire tests, not just single questions. And finally we explore the psychometric implications of edge aversion.

Where Do People Place the Correct Answer When Constructing a Single Multiple-Choice Question? A Priori Hypotheses and Normative Considerations

Imagine someone contributing a single question for a multiple-choice test. Where should the correct answer be placed? "Wherever" or "At random" seem to be pretty reasonable answers. Where in fact do people place the correct answer? A

priori, several possibilities come to mind. The following hypotheses all pertain to a single isolated question with k answer options, not to a set of questions:

1. Anchoring. The order of writing the answers follows the order in which they occur to the writer. Because a question and its correct answer seem to be a natural point of departure for a multiple-choice item, the writer might tend to start by writing them down, and only then attempt to construct distractors. If so, there would be a preponderance of correct answers (i.e., more than $1/k$) in the first position. If writing order mimics invention order, and the writer has a hard time finding that last suitable distractor, then perhaps once found, it will occupy the last remaining slot, leading to a preponderance of incorrect answers (i.e., more than $(k–1)/k$) in the last position.

For numerical questions, anchoring might lead to a different order, due to the tendency to bracket the correct number—the anchor—by distractors both larger and smaller (Inbar, 2002). Because numerical answer options are often re-ordered monotonically (a recommended rule of thumb—see Haladyna & Downing, 1989, rule 25), thus disguising the original order in which they came to mind, the correct answer would be found in central positions more often than $(k–2)/k$.

2. Narration style. Perhaps the writer wishes to lead up to the correct answer with a kind of tension build-up: "Is the answer A? No, it is not. Is it B? No, it is not. . . . Is it, perchance, E? Indeed it is!" This scheme is an effective rhetorical device, often found in children's stories. In addition, perhaps putting the correct answer last seems to distance it from the test taker most, serving to better hide it. Of course, when the correct answer happens to be a meta-answer, such as "None of the above" or "All of the above" (both ill-advised answer options, see Haladyna & Downing, 1989, rules 29–30), the last position is the most natural, if not the only, possibility. All these possibilities would lead to a preponderance of correct answers in the last position.

3. Compromise. Perhaps question writers don't really care where they put the answer and gravitate toward the center as a kind of compromise of the set of possible positions. The center may also feel like a less conspicuous place to "hide" the correct answer (see evidence in the next section). If so, there would be a preponderance of correct answers in central positions. This article will show that this is in fact the observed bias.

4. Randomization. An absence of any systematic bias within a person, or an absence of any single dominating mechanism or strategy for placing correct answers across people, may lead to pseudo-randomness—no systematic preponderance of correct answers in any particular position. Perhaps this is all that psychometric theory intends or requires in its implicit assumption that the positioning of correct answers is randomized. But as to individual randomization, it is well known that people are quite incapable of simulating a random device in their heads and can only randomize by actually resorting to an external random device, such as dice (e.g., Bar-Hillel & Wagenaar, 1991). Nonetheless, it is hard to overstate how normatively compelling randomization is in placing correct answers.

Having put forth some hypotheses about possible positional tendencies, the next section will survey and report empirical evidence regarding where people actually place correct answers when constructing multiple-choice questions.

Where People Place the Correct Answer When Constructing a Single Multiple-Choice Question: Empirical Evidence

This question cannot be answered merely by observing published tests because when constructing an entire test, test makers often move the correct answer from the position in which it appeared originally to obtain a balanced answer key. The closest thing to a study of single questions that we found in the literature was reported by Berg and Rapaport (1954). Four hundred students in a course given by Rapaport were required "to prepare four multiple-choice items each week on the lecture and reading materials" and write each question on a separate card, "no directions concerning answer placement were given" (p. 477). In 1,000 consecutive cards from the first assignments, correct answers were distributed over the four positions as follows: 54, 237, 571, and 138, totaling 80% in the middle.

In collecting our own data, we asked 190 widely assorted people to write a single four-choice question on the topic of their choice. Of these, 125 received the task embedded within a questionnaire alongside other, unrelated tasks. Their written instructions were to invent a question to which they knew the answer, invent three distractors, and write all down in the provided slots. They were told to avoid answers such as "All of the above" or "None of the above" or questions that had numerical answers. Answer positions were labeled A, B, C, and D—but left empty. The respondents were a convenience sample consisting of native Hebrew speakers with at least a high-school education, recruited one by one in arbitrary fashion. Their mean age was 29, and half of them were male. Correct answers were distributed over positions as follows: 31, 40, 41, and 13. Altogether, nearly 70% of the answers were placed in central positions. Only 50% would have been expected by chance ($p < .0001$).

The other 65 people, all acquaintances and colleagues of the authors, were approached informally: either students and faculty in The Hebrew University's Psychology Department or personal friends. The academics, experienced in writing multiple-choice tests, were instructed to write a four-choice question on any topic they wished, avoiding answers of "All answers are true" or "No answers are true." The personal friends were told, "Write a question for [the popular TV program] 'Who Wants to be a Millionaire.'" They jotted down their responses on small pieces of paper. No explanation was offered. Because they did not differ on the dependent variable, we report on all together. The distribution of correct answers over positions was 7, 21, 30, and 7—nearly 80% in central positions ($p < .0001$).

None of the debriefed acquaintances suspected what the point of the little exercise was. They were quite surprised to hear its true purpose—and doubly surprised at the results. They admitted no insight as to why they had placed their answers where they did and indeed seemed to have none. When inventing their questions, position was the last thing on their mind, and correct answers were positioned with little if any deliberation.

Where Do People Seek the Answer to a Single Multiple-Choice Question?

As mentioned before, whereas test makers have complete freedom to place the correct answer wherever they choose, test takers merely want to reproduce the test makers' choices. The natural way to choose among the multiple offered answers is

by content, not position. Only when guessing might a test taker consider position. Hence it is harder to determine where people seek the correct answer to a multiple-choice question than to determine where they hide it.

To ensure guessing, a class of 127 undergraduate psychology students was asked—in the context of a longer questionnaire—to imagine that they were taking a multiple-choice test with four options. Two questions were presented, but the content of the answers was missing and only their positions given, as shown below. The questions, in order, were

What is the capital of Norway?
A B C D

What is the capital of The Netherlands?
A B C D

Obviously, the respondents could not actually answer the question, but they were requested to nonetheless guess where it was. Sixty nine responded to both questions, and 58 others found the answer to the first question already circled (A and D were circled for 15 respondents each, and B and C were circled for 14 respondents each), and had to respond only to the second question. The distribution of the 69 position choices in the Norway question was 7, 33, 28, and 1. In the subsequent question, it was 11, 21, 26, and 11 if the first question had been self-answered ($N = 69$), and 12, 20, 22, and 4 if the first question had been pre-answered ($N = 58$). In toto, the percentage of instances each answer position was chosen over the 196 choices (69 + 69 + 58) made in both conditions and both questions was 15%, 38%, 39%, and 8%-almost 80% middle choices ($p < .0001$).

Berg and Rapaport (1954) report similar results when not only the answer contents were lacking but the question, too. In other words, they gave 374 students what they called an "imaginary questionnaire" (Table 2, p. 478) consisting of nine various kinds of forced-choice "imaginary questions." In their Question 2 the answers were labeled 1, 2, 3, and 4, and in Question 8, they were labeled A, B, C, and D. In an inverted form given to 203 other students, the labels were 4, 3, 2, and 1 and D, C, B, and A, respectively. The results were [1–14, 2–41, 3–92, 4–24], [4–31, 3–89, 2–60, 1–23], [A–31, B–64, C–47, D–29], and [D–27, C–55, B–78, A–43], respectively, for a total of 70% middle choices ($p < .0001$).

Test takers seem to be unaware of their tendency to guess middle positions. We asked 40 Israeli students "What does the Japanese word *toguchi* mean?" The possible answers and their response frequencies were the following: A. Door (8), B. Window (23), C. Wall (3), and D. Floor (6)—65% middle answers ($p < .05$). Only two students explained their choice by mentioning position explicitly ("I just gave the first answer" and "C just grabbed me").

Our final piece of evidence on guessing in a single question comes from a real test. The Psychometric Entrance Test (PET) is a timed four-choice test designed by Israel's National Institute for Testing and Evaluation (NITE). The PET, like the SAT, which it resembles, is used in the selection of students for admissions purposes by Israeli universities and measures various scholastic abilities. It consists of two quantitative sections, two verbal sections, and two English sections. The population of the PET test takers is nearly gender balanced (54% females), consisting of young (over 90% are under 26) high-school graduates.

In the PET's quantitative sections, questions are ordered from easy to difficult. Because test takers often run out of time toward the later questions, some appear to

TABLE 1
Percentage of middle choices in long runs

Run length	*N* of tail runs	% in middle
4	1266	64
5	534	78
6	184	83
7	97	91
8	49	84
9	24	83
≥10	30	87

give up toward the end of the test, foregoing even the appearance of attempting to discern the correct answer in favor of guessing in a speedy but completely arbitrary way. A rare strategy (less than 1% of all test takers), but an easily identifiable one, is to choose a single position and mark just it from some point in the test on. Arguably some of these runs may genuinely reflect the test taker's best guess, but the longer the run, the less likely that is.

The responses of real test takers to five different versions of the quantitative subtest of PET, each consisting of two 25-item sections, were analyzed. All in all, 35,560 examinees took these two sections, yielding 71,120 25-answer sequences. Table 1 shows tail-runs (i.e., runs of identical responses ending in the final item) of various lengths and the percentage thereof in which a central position was chosen. By and large, the longer the run, the higher the proportion of middle positions. We interpret this to mean that the larger the probability that a run is really a "give up" and guess (or the higher the percentage of pure guessers constituting the run), the larger the edge aversion, till it matches or surpasses the magnitude of edge aversion in guessing single questions. Note that though these are multiple responses, they still represent a single choice—the one starting the run.

We have shown that people writing isolated four-choice questions hide the correct answer in the two middle positions about 70% of the time, and people guessing an isolated four-choice question seek it in the middle about 75% to 80% of the time. Is it possible that the guessers favor the middle simply because they believe this mimics what the test makers are doing? We consider this possibility unlikely for several reasons.

First, test writers' edge aversion, though it seems to be part of multiple-choice testing lore (see the opening quotation), is not explicitly acknowledged by any test-coaching book or course for SAT preparation that we encountered in a casual survey of such books. Perhaps this reflects the implicit assumption that professional tests have already corrected it, which is by and large true (see the following section). Second, people encounter most multiple-choice questions within sequences (i.e., entire tests), rather than in isolation. In entire tests edge aversion is much diluted because entire tests often correct for the single-question middle bias. Hence, test takers would have a much-reduced opportunity to encounter it. Third, the normative response to a middle bias when guessing is to choose a middle position all of the time, not just most of the time. Last but not least, in the following

section we show that a tendency toward the middle—or away from the edges—exists in contexts that have nothing to do with tests or with experience. Edge aversion in the context of tests may be no more than another manifestation of edge aversion in its general form.

Edge Aversion

In a four-choice test, it is hard to say a preponderance of guesses in the middle reflects an attraction to the middle or to an aversion to the edges. But a similar bias has been observed in many other tasks, in some of which it is clearly edge aversion that underlies the bias (see 2, 3, and 5 following).

1. The closest task to a multiple-choice test was studied by Rubinstein, Tversky, and Heller (1996). Respondents were told to "hide a treasure" in one of four places laid out in a row, where others would then seek it by getting a single opportunity to observe the content of a single hiding place. "Hiders" won if the treasure were not found, whereas "seekers" won if it were. Respondents in both roles favored middle positions (thereby undermining any notion that the bias was somehow related to strategic advantage). The authors seem to have shared our intuition regarding what drives this bias, talking about "players' tendency to avoid the endpoints" (p. 399).

2. Falk (1975) asked respondents to mark 10 cells in a 10 x 10 matrix "as if these cells were chosen by blind lottery." The cells on the border of the grid were heavily avoided. Indeed, the median (and modal) number of edges that the marked cells shared with the border was two—half the number expected from randomly marked cells. However, within the array's 8 x 8 interior, the middle was not systematically preferred, supporting the notion that the bias toward the middle is really a bias away from the edges.

3. In a kind of extension of both Falk's (1975) and Rubinstein et al.'s (1996) work, Ayton and Falk (1995) asked respondents to mark three cells in a 5 x 5 matrix under a very wide variety of instructions. Under any instructions that evoked, explicitly or implicitly, a hide and seek context, edges were avoided—but so was the exact middle cell. Under instructions that encouraged going for the salient cells, the four corners and the exact middle were the favorites. Excluding the exact middle cell and the four corner cells, interior cells were more popular than border cells under all instructions—even though in 5 x 5 matrices there are more border cells than interior cells (whether or not the corners and the middle are excluded).

4. Five-choice tests such as the SAT allow a distinction between middle attraction and edge aversion. In five-choice tests, while positions A and E are the least popular, position C—the precise middle—is not the most popular (see Table 2), suggesting that it is not so much attraction to the middle as aversion to the extremes that underlies this middle bias.

5. Ullman-Margalit and Morgenbesser (1977) made the distinction between choosing and picking. Whereas the former "is determined by the differences in one's preferences," picking occurs "where one is strictly indifferent" (p. 757). A prototypical picking situation arises when one selects, say, a coffee jar from an array of identical such jars in the supermarket. In our world, otherwise identical objects (e.g., coffee jars) necessarily occupy different places in space-time (e.g., supermarket shelves). In a study called "Choices from identical options," Christen-