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

TABLE 2
Percentage of correct answers by positions in various answer keys

Test	Number of questions	A	B	C	D	E	% in middle
4-choice test							
PET pilot[a] (1997-1998)	8905	25	26	25	24	-	51[b]
10 Operational PET tests[a] (1997-1998)	1640	25	24	23	27	-	48[b,c]
Yoel (1999)	2312	24	28	27	21	-	55[b]
Offir & Dinari (1998)	256	20	27	29	24	-	56[b]
Kiddum (1995)	1091	24	26	26	24	-	52
Open University (1998)	258	27	27	25	21	-	52
Gibb (1964)	70	24	34	21	20[c]	-	56[b,c]
Trivia (1999)	150	23	27	27	23	-	53[c]
SAT (Claman, 1997)	150	29	23	23	25	-	47[c]
5-choice test							
SAT (Claman, 1997)	1130	19	20	22	21	19	63[b]
MPT[a] (1988-1999)	1440	18	22	21	21	18	64[b]
INEPE (1998)	432	18	25	21	19	18	64[b,c]
GMAT (GMAC, 1992)	402	17	19	23	22	19	64

Note. MPT = Mathematics Proficiency Test; PET = Psychometric Entrance Test; INEPE = Instituto Nacional de Estudos e Pesquisas Educacionais.

[a] Answer keys courtesy of the National Institute for Testing and Evaluation, Israel.

[b] Significantly different than expected (50% in 4-choice tests; 60% in 5-choice tests), $p < .05$.

[c] Total is not the exact sum of its constituents due to rounding.

feld (1995) found that "Whether people were choosing a product from a grocery shelf, deciding which bathroom stall to use, or marking a box on a questionnaire, they avoided the ends and tended to make their selection from the middle. For example, when there were four rows of a product in the supermarket, . . . 71% [of the purchases] were from the middle two" (p. 50). Similarly, middle toilet stalls were chosen 60% of the time. Thus, though the tendency of one's fellow folk to select products from the middle of a grocery shelf makes it strategically advantageous for oneself to do the same (because products there will tend to be replaced more often, hence fresher), whereas the tendency of one's fellow folk to use central bathroom stalls makes it strategically disadvantageous for oneself to do the same (because those stalls will tend to be dirtier and run out of toilet paper more often), middle bias is evident in both. Edge aversion does not need to be strategically advantageous in order to occur.

6. In a task devoid of any strategic content, Kubovy and Psotka (1976) asked subjects to "give the first number between 0 and 9 that came to mind" (p. 291). In their own study, as well as a number of other studies they surveyed, the numbers 0, 1, 9, and 2 (ordered by scarcity) were chosen less frequently than 7% each (compared with 10% by random choice). The favorite numbers were 7

(nearly 30%), followed by 3 (nearly 15%). These results indicate edge aversion rather than middle favoring: 4, 5, and 6 were not as popular.

Although in the opening section we derived the prediction of a middle bias under a heading of "compromise," we have no direct evidence for that notion. The middle bias in tests is probably just another manifestation of the general phenomenon of edge aversion, which in spite of some attempts (e.g., Shaw, Bergen, Brown, & Gallagher, 2000), is a phenomenon still in search of an explanation. We will use the terms *middle bias* and *edge aversion* interchangeably, without committing to any underlying causal mechanism.

Do Tests Answer Keys Exhibit Middle Bias?

In an informal, non-test context, we have shown that whether people are hiding an answer to a single four-choice question or seeking it, middle positions are favored at a ratio of about 3 or 4 to 1. Does the same happen when people are either writing or answering an entire test—namely a long sequence of multiple-choice problems? This section addresses this issue from the perspective of test makers, and the following section from that of test takers.

The dominant, if not publicly acknowledged policy regarding answer keys, favored by professional test makers, is known as key balancing. Notably, this is the policy at one of the world's largest testing organizations, the Educational Testing Service (ETS, makers of the SAT and the GRE, among others). As a result, any edge aversion that might have existed in the "raw" answer key is corrected and all but disappears in the answer key's final version.

Nonetheless, traces are still left. Table 2 shows the proportion of correct answers that were placed in each position in a number of tests. This collection of tests, though arbitrary, was nonetheless not selected from any larger set we considered. Some of the tests included (SAT, GMAT, PET) are produced by the world's most professional organizations. NITE gave us access to answer keys of some tests they developed (e.g., 8,905 pilot PET questions; 10 operational PET exams; the Mathematics Proficiency Test (MPT)—a mathematics exam for student selection at Israel's Technion. The College Board's publications provided us with 10 SAT keys (Claman, 1997) and GMAT keys (Graduate Management Admission Council, 1992). Correct answers to trivia questions were taken from the website www.triviawars.com. The other answer keys appeared in published sources, as cited. Of the 13 sources we considered, 11 produced tests with an over-preponderance of middle answers ($p = .01$, sign test).

An "anchor and adjust" procedure could explain why the middle bias has not been completely eradicated. Key balancing may be done iteratively, stopping when the proportions are roughly—though not necessarily precisely—equal. But starting, as it does, with a preponderance of answers in the center, chances are that the process will halt too soon, when this preponderance, though considerably diminished, is still there.

We found three previous surveys of a similar kind, all showing the same traces of middle bias in operational tests. McNamara and Weitzman (1945) analyzed nearly 5,000 five-choice questions and 4,000 four-choice questions taken from operational tests, and found correct answers placed in center positions in 62% of the former and

TABLE 3
Percentage of sections with a given majority of correct answers in middle or extreme positions

Majority (out of 25)	Observed % of sections	Expected % of sections
13	55	31
14	34	27
15	8	19
16	2	12
17	1	6
18	0	3
19	0	1
20 to 25	0	0

51% of the latter. Metfessel and Sax's (1958) summary of their harder-to-quantify results was, "There is a tendency for the authors of multiple-choice tests to place the correct answer at the center" (p. 788). Finally, Mentzer (1982) presented a table much like ours, reporting the distribution of correct answers over positions in a four-choice test, and in 30 of 35 tests, the answer was more often in a central position ($p < .0002$, sign test).

Ironically, while the mean difference between the observed percentage of middle responses and the expected percentage is positive, hence too large, the absolute difference is too small. In other words, though the proportion of middle answers is significantly higher than expected by chance (52% in the four-choice tests; 64% in the five-choice tests), the variability around this proportion is too small: There are not enough sections in which there is a major imbalance between, say, middle positions and edge positions.

Table 3 shows the percentage, among 109 four-choice PET sections,[2] in which a given majority of correct answers was found, either in the middle or in the extreme positions. Because these sections are 25 items long, a majority can range from 13 to 25. The table also shows the expected percentage under random positioning of the correct answer. Balanced keys (i.e., 13 or 14 correct answers in the middle or the extreme positions) are over-represented in the actual tests, and unbalanced keys (15 and up) are under-represented.

Because SAT sections vary in length, we cannot produce a similar table for the SAT. So we measured over-balancing in SAT keys as follows: For each of the SAT's six sections, we computed the absolute difference between the expected number of correct answers in the extreme positions (assuming random choice) and the observed number. For example, in a section of 25 questions, if eight correct answers are in position A or E, rather than the expected 10, then the absolute difference is two. The sums of these absolute differences (SADs) over all six sections of 10 real SAT tests (Claman, 1997) are presented in the first row of Table 4. We next conducted a Monte Carlo simulation on 100,000 randomly generated SAT answer keys (i.e., correct answers were randomly assigned a position based on $p = .20$) and computed the SADs for each. This yielded a distribution of expected SADs with a mean of 10.5 ($SD = 3.4$), and a median of 10.3. The second row of

TABLE 4

Observed SADs[1] in 10 real SAT test keys and their percentile rank in the expected distribution under randomization

SAD	4.7	5.1	5.7	5.7	6.7	7.3	7.9	8.7	9.9	12.7
Percentile Rank	4	4	7	7	13	17	25	32	47	75

[1] Sum of Absolute Difference.

Table 4 shows the percentile rank of the observed SADs in the distribution of SADs obtained by the Monte Carlo. Note that all but one of the 10 observed SADs has a percentile rank less than 50 ($p = .01$, sign test), and the median percentile rank is 15.

Merely for ease of comparison, we performed the same analysis for the PET, using the 10 operational PET tests mentioned previously. The median SAD there was also in the 15th percentile, and all 10 tests had an SAD percentile rank lower than 50.

In summary, in key-balanced tests, the distributions of correct responses over positions are both overly balanced (too close to uniformity) and not balanced enough (still middle-biased). The former is a deliberate and intended consequence of key balancing, to be expected whenever this policy is practiced. The latter is probably unintended and might result from an iterative anchor and adjust balancing process. It is not a necessary feature of key balancing but evidently happens both in the PET and the SAT.

Are Examinees in Real Tests Middle Biased?

Because test makers produce balanced keys, test takers, insofar as they mimic them, should also produce balanced keys. Hence, test takers have an opportunity to exhibit their own edge aversion only with respect to items on which they have imperfect knowledge. Besides the fact that the scope for examinees' edge aversion in a full-length test is thus smaller than the full length of the test, another factor that might attenuate edge aversion in a full test, as compared to a single question, is people's propensity to give representative predictions under uncertainty (Kahneman & Tversky, 1972). In a full test, test takers who believe that the correct answer key is balanced may wish to produce a representative, balanced answer sequence on their own, even if they cannot quite reproduce the correct one. In other words, they may believe that they increase their chances of answering correctly if they produce an answer sequence whose properties reproduce those of the correct answer key.

These two robust, documented tendencies—edge aversion and representativeness—conflict when guessing in a sequence of questions because the first leads to unbalanced guessing (middle bias) and the latter to balanced guessing. So we analyzed real test results, as well as experimental data, to see whether and to what extent test takers nonetheless exhibit edge aversion in real tests.

The most straightforward data on this issue is the distribution of chosen answers over answer positions in the answer sequences of real test takers. Cronbach (1950) was quite confident there would not be any "response set" in his terminology, and others have concurred in this expectation. But when considered empirically, "The

TABLE 5
Percentage of examinee answers in each position

Position	A	B	C	D	Middle (B + C)
All answers	23	26	27	23	53
Erroneous	22	27	28	23	55
Correct key	25	26	25	24	51

few studies on positional bias . . . are inconclusive because of methodological and conceptual problems" (Fagley, 1987, p. 95) and perhaps also because such biases as were found were typically weak. Some authors have found a preference for early responses (Clark, 1956; Gustav, 1963), some for middle positions (Rapaport & Berg, 1955), and some no positional preferences at all (Marcus, 1963) or mixed ones (McNamara & Weitzman, 1945).

We looked at the answers of the examinees who took the 8,905 questions piloted by the NITE in 1997–1998 (about 300 per question). Table 5 shows the percentage of all middle answers, 53%, as well as the percentage of erroneous middle answers only (errors constituted just over a third of the total), 55%. The examinees, we see, produced answer sequences with a middle bias only slightly higher than observed in the correct answer key, 51% (from Table 2).

Additional evidence comes from data collected, for different purposes, by Bereby-Meyer, Meyer, and Budescu (2000). Among 34 ordinary questions, they inserted 6 nonsense questions—namely, questions to which there are no correct answers (e.g., "Where does the Lagundo tribe live? A. West Africa B. New Guinea C. Ecuador D. Indonesia"—there is no Lagundo tribe). The 144 respondents preferred middle options to extreme ones 53% of the time.

Data from distributions of chosen answers over answer positions provide us with limited information about the amount of edge aversion of real test takers. In the first place, the total amount of answers include many correct answers that result from true knowledge. Second, even wrong answers include many answers that result from incorrect knowledge and not from guessing. In the next section we present data from an experimental manipulation that reveals the edge aversion of guessing examinees.

Switching the Position of Answer Options

If guessing examinees' attraction to different answer options is influenced by answer position, changing the positions of the answers should reveal this bias. Specifically, suppose one group of examinees receives questions with the answers in positions A, B, C, and D, and another group gets the same questions with the same answers but reordered into positions B, A, D, and C. So answers that were presented in the middle in one version appear on the edges in the other—and vice versa. If guessing test takers are edge averse, more examinees would choose options B and C in the first presentation than in the second, while A and D would be more popular in the second version than in the first. We carried out just such an

TABLE 6
Difference between, and ratio of, the percentage of examinees who chose a certain (wrong) answer when it was in a middle position versus an extreme one

Original position of distractor	Number of questions of this kind	% choice in extreme position	% choice in middle position	Mean difference (and its *SD*)	Median ratio
A	116	9.1	12.1	3.0 (4.0)	1.32
B	116	10.0	12.3	2.3 (3.1)	1.18
C	123	10.3	13.3	3.0 (3.6)	1.24
D	128	10.1	13.1	3.0 (3.6)	1.35

experiment on six assorted PET sections, containing a total of 161 questions. Each section was answered by a different group of at least 220 examinees, for a total of about 4,000. The examinees who participated in this experiment did so while taking the PET, which hides a pilot section among the operational sections.

Two dependent variables were used—the difference between the percentage of examinees choosing an option when it was presented in a middle position versus in an extreme position and the ratio of these percentages. We focused on wrong answers on the general assumption that these are mostly guessed answers. On average, the three wrong options attracted about one third of the examinees or about 11% for each one. Table 6 shows that the mean difference in the percentage of examinees choosing a certain wrong option when it was presented in a middle position was higher by almost 3 percentage points than when it was presented in an extreme position ($t_{482df} = 17.5, p < .0001$). The median ratio between the proportion of examinees choosing a wrong option when it was presented in a middle position versus an extreme position was 1.29. In other words, the 3% difference means that nearly 30% more examinees chose a wrong option when it was presented in a middle position.

Estimating the Middle-Bias Magnitude in Real Tests

The following knowledge model allows an estimation of the magnitude of the middle bias. Let us partition the set of examinees answering a given question into those who know the answer, labeled K, and those who do not and consequently must guess it, labeled G. The examinees can also be partitioned into those who choose one of the middle options as their answer, labeled M, and those who choose one of the extreme options, labeled E. Clearly, E is the complement of M and, by assumption, G is the complement of K. Finally, we label the examinees that choose the correct answer C.

We label by *M* or *E* the event that the correct answer is, respectively, in a middle position (B or C) or in an extreme position (A or D). *M* differs from M (and *E* from E) in that the former denotes the true position of an answer, and the latter denotes the test taker's choice. The assumption of standard psychometric models is that a guessing examinee has no preferences among positions, hence the probability of guessing each position is 1/4. Our model makes the weaker assumption that guessing examinees have no preference between the two central positions or between the two extreme positions, but it does not require that P(M) = P(E). In addition we make the natural assumption that knowing an answer is independent of its position, so that P(G) (or P[K]) are independent of P(*M*) (or P[*E*]). Hence,

$$P(C|E) = P(K) \cdot 1 + P(G) \cdot P(E|G) / 2 = [1 - P(G)] + P(G) \cdot [1 - P(M|G) / 2 \quad (1)$$

$$P(C|M) = P(K) \cdot 1 + P(G) \cdot P(M|G) / 2 = [1 - P(G)] + P(G) \cdot P(M|G) / 2 \quad (2)$$

Because P(C|*E*) and P(C|*M*) are observables, the two equations include only two unknowns. The first, P(G), is related to the question difficulty. The second, P(M|G), is the magnitude of the bias to the middle, assumed to be constant across questions. In the terminology of hypothesis testing, the null hypothesis suggested by standard psychometric models claims that P(M|G) in a four-choice question equals 1/2, whereas our research hypothesis is that it is larger than 1/2.

Isolating P(M|G) from Equations 1 and 2 we obtain,

$$P(M|G) = [1 + P(C|M) - 2P(C|E)] / [2 - P(C|M) - P(C|E)]. \quad (3)$$

For example, when P(C|*M*) is .67 and P(C|*E*)] is .64, as we found, then P(M|G) is 0.56. But we used 3 to estimate P(M|G) separately for each of the 161 questions in the switching experiment. The resulting estimate was significant at 0.57 ($t_{160} = 6.9$, $p < .0001$). This is actually an underestimation because the model simplistically ignores examinees whose errors are not due to guessing (they may be due, e.g., to an erroneous belief in the correctness of some option). Whenever this happens, P(G) will actually be smaller than 1 − P(K) and then, to obtain a given difference between P(C|*M*) and P(C|*E*), the true value of P(M|G) must be larger than the estimate based on Equation 3. So we found that at least 57% of the examinees who do not know the answer to a question choose to guess a middle position.

Why is the tendency to guess a middle position in a test-embedded item so much weaker than the tendency to guess a middle position in a single question (nearly 60% versus nearly 80%, respectively)? This might be due to the fact that edge aversion is inconsistent, over more than a few items, with a balanced key. A preference for middle positions and a preference for a balanced answer sequence conflict, of course. Perhaps the 6:4 ratio is a kind of compromise between these two tendencies—the 4:1 ratio in guessing an isolated item, and the 1:1 ratio of a perfectly balanced key. Test takers' inability to overcome the middle bias completely and achieve a balanced key may be due to the attentional effort such balancing requires—in contrast to the effortless middle bias.

Psychometric Consequences

Edge aversion is a bias with which psychometrics must contend because it has implications for psychometric indices, such as item difficulty and item discrimination. Insofar as guessing examinees have a preference for middle options, questions with middle correct answers will be easier and less discriminating than these same questions with the same correct answers placed at the edges because more low-ability examinees will choose the correct answer in the former case.

TABLE 7

Effect on item indices of difficulty and discriminability of moving the correct answer from a middle position to an extreme one

Statistic	*N* of items	*M*	*SD*	Effect size	*t* value
Difficulty (drop in % correct)	161	3.3%	5.1	0.22	8.1*
Discriminability (increase in the biserial)	161	0.05	0.1	0.32	5.5*

* Difference for the one-tail t test is significant ($p < .0001$).

Position Effects on Item Difficulty and Discriminability: Evidence From the Switching Experiment

Table 7 presents the difference in item indices between the two answer presentations (once in a middle position and once in an extreme position) of the questions in the switching experiment.

Note that the effect of position (middle versus extreme) on percentage correct responses, 3.3%, is even larger than its effect on incorrect responses (Table 6, column 5). However, these averages do not present the full picture. Clearly, the magnitude of the position effect depends on the difficulty of the question: the harder the question, the larger the percentage of guessers. Correspondingly, the position effect can be expected to be larger, too. Indeed, the regression models for predicting both indices—difficulty (measured by percentage correct) and discriminability (measured by the biserial correlation)—from the mean difficulty of the item (over the two presentations) are significant ($p < .01$). Table 8 presents the predicted values of the position effects.

Position Effects on Item Difficulty and Discriminability: Evidence From the PET Pilot Questions

In addition to the switching experiment, we estimated the position effect on item difficulty and item discriminability using data from real examinees in a natural testing situation. All 8,905 PET questions piloted during 1997–1998 were analyzed (piloted questions were chosen rather than operational ones because the latter had already undergone selection based on their difficulty and discriminability, as estimated without regard to the position of the correct answer). Questions with middle correct answers were indeed easier (64% correct) than questions with extreme correct answers (61% correct). The difference, 3%, resembles the mean difference in the controlled-switching experiment, 3.3%, even though here we do not compare identical questions, as we did there.

To see the dependency of this effect upon question difficulty, we considered only the quantitative questions because they are conventionally placed in ascending order of difficulty. Table 9 presents the mean effect for the quantitative questions by the position of the question in the section—a proxy for question difficulty.

The gain in percentage correct due to the middle bias rises from 1% to 7% as difficulty increases—an even wider range than predicted by the regression model

TABLE 8
Predicted effect of moving the correct answer from a middle to an extreme position on items' difficulty and discriminability, as a function of question difficulty

Difficulty (% correct)	90	80	70	60	50	40	30
Difficulty (drop in % correct)	1.3	2.1	2.9	3.7	4.6	5.4	6.2
Increase in biserial	.01	.03	.04	.05	.07	.08	.09

TABLE 9
Percentage of correct answers in middle-keyed and in extreme-keyed items, by item's ordinal position in the section

Ordinal position	1–6	7–12	13–18	19–25	All
Middle keyed	72	65	60	50	62
Extreme keyed	71	62	54	43	57
Difference	1	3	6	7	5

(Table 8 shows a 3% to 5% effect for a similar range of difficulties, namely between 40% and 70% correct).

In order for a question to become operational it obviously needs to be sufficiently discriminating. At the NITE, a "good" question is one with a biserial value for the correct answer of at least 0.3 and point biserial values for all distractors of, at most, −0.05. It seems odd that the quality of a question would depend on the position of the correct answer. Yet Table 10, in which the questions were sorted into quartiles by difficulty, shows the difference in percentage of good questions among those with middle versus extreme correct answers. For both types, it presents the percentage of successful questions, the percentage of inadequate questions (i.e., biserial value lower than 0.3 for the correct answer or point biserial value higher than −0.05 for at least one distractor), and the mean biserial values. Evidently, the position of the correct answer affects the various discrimination values, and this effect becomes very large for difficult questions.

Position Effects on Item Response Theory-Based Item Indices: Evidence From Operational PET Questions

Our last analysis examines position effects on item response theory (IRT)-based item indices. At the NITE, IRT parameters of PET items are estimated only for operational items because piloted items are not administered to enough examinees (typically only 200–300), whereas operational items are typically administered to several thousand examinees.[3] Because operational items are selected on the basis of their discrimination, position effects will be smaller than in piloted items.

We compared item parameters of all 4,533 operational PET items from the years 1997–1999. Table 11 shows means of the *a*, *b*, and *c* parameters for middle-keyed and extreme-keyed questions. For all three parameters we observe the expected position effects: Middle-keyed questions are less discriminating (lower *a* value), easier (lower *b* value), and have a higher pseudo-guessing parameter value *c* than

TABLE 10
Percentage of "good" questions, and mean biserial, by position of correct answer and difficulty quartile

	Middle correct position					Extreme correct position				
Quantile	Q1	Q2	Q3	Q4	All	Q1	Q2	Q3	Q4	All
Sufficient discrimination	85	83	70	34	70	86	87	78	51	75
Biserial ≥ .3	93	93	83	58	83	95	96	90	75	88
Point-biserial of distractors ≤ −.05	87	86	74	38	73	88	90	82	53	77
Mean biserial	.52	.49	.43	.32	.45	.55	.53	.49	.39	.48

TABLE 11
Means (and standard deviations) of item response theory item parameters (based on Psychometric Entrance Test questions)

Position	*N* of items	a	b	c
Middle	2183	.83 (.35)	−.19 (1.04)	.22 (.08)
Extreme	2350	.86 (.34)	−.01 (0.98)	.18 (.07)
Standardized difference	4533	.09	.17	.49

extreme-keyed questions—all effects are statistically significant ($p \leq .001$). The standardized differences (i.e., the mean differences divided by the total *SD*) range from about a 10th of the standard deviation for *a* to about half of the standard deviation for *c*.

Position Effects on Successful Guessing

One component of question difficulty is the probability of guessing it correctly. The shared preference of examiners and guessing examinees for placing correct answers in middle positions brings about a positive correlation between their choices. This positive correlation is, of course, advantageous to examinees, enhancing their probability of successful guessing. If the distribution of percentage of guessing examinees over positions is (P_1, P_2, P_3, P_4) and that of percentage of examiners positioning correct answers is (Q_1, Q_2, Q_3, Q_4), then the probability of a match between the choices—a match leading to a successful guess—is $P_1Q_1 + P_2Q_2 + P_3Q_3 + P_4Q_4$, assuming independence between examiners and examinees. In professional tests, where the middle bias is typically about 52% in the key and about 55% for examinees, the match probability in a four-choice test item is negligible. But if the bias is as high as 80%, as it is for single isolated items, then the probability of a match can go up to 40%.

If test makers randomize, thereby producing a bias-free answer key (on average), it does not matter whether the test takers are biased or not—their match probability will be 25%. In that respect, if answer keys were randomized, it would not matter whether examinees were middle biased or not. But examinees' middle bias does matter regarding the following issue.

How to Obtain a Randomized Operational Test

Suppose a set of questions is keyed at random, and then the questions are assessed for discriminability. Because the examinees are middle-biased, middle-keyed items will—other things equal—be less discriminating than edge-keyed items and thus have a smaller chance of "passing" the minimal requirements for a good item. Consequently, a bank of piloted key-randomized items will yield an operational question bank with a preponderance of edge-keyed items. For difficult questions, the effect can be very large. For example, in Table 10, 51% of the extreme-keyed questions in the fourth quantile have sufficient discrimination, but only 34% of the middle-keyed questions in Q4 have sufficient discrimination. Hence, if one starts, for example, with 100 edge-keyed difficult items and 100 middle-keyed difficult items, one will end up with 85 operational difficult items (51 + 34), of which 60% will be edge-keyed. If items are selected for tests from this edge-biased bank at random (namely, with disregard for the bias), the entire test might actually end up being edge biased rather than middle biased. Indeed, we suspect this is the reason why the 10 operational PET tests recorded in Table 2 had only 48% middle answers: They were selected from an almost perfectly balanced pre-selection pool (51% middle answers in the PET pilot, see Table 2).[4]

To obtain a balanced operational test, items should be stratified by answer position (namely, questions should be separated into those that are A-keyed, B-keyed, etc.), and which group should be sampled for each ordinal position within the test sequence should be determined at random. The same recommendation applies to adaptive testing. In addition, the difference between an edge-keyed and middle-keyed question with the same distractors allows one to play around with correct answer position in the following manner: If a middle-keyed item is only borderline good, it might be salvaged by piloting it again with the correct answer in an extreme position. On the other hand, if an edge-keyed item is comfortably good, it could remain usable even if the answer were moved to a middle position.

Conclusion

Cronbach (1946) drew a list of about half a dozen "response sets," defining them as "any tendency causing . . . different responses to test items than . . . had the same content been presented in a different form" (p. 491). Cronbach's list included no position-related response sets. Ironically, among his suggestions "for eliminating the effects of response sets upon test validity," the first is, "The multiple-choice form, which appears free from response sets" (p. 492). The present article focused on the middle bias, which is not only a hitherto unacknowledged response set of test takers but also, according to informal test-wise lore, a bias of test makers. Our reported analyses focused on tests produced by professional test makers, among whom middle bias is all but removed by the common and careful practice of key balancing (although a slight trace remains, see Table 2). Test takers, however, are a different story. Though they, too, exhibit a tendency to balance their answer keys (most obviously manifested in their over-alternation when guessing), their guesses nonetheless contain a slight preponderance (roughly 10%) of middle responses. Unlike some response sets, middle-bias is not detrimental to test takers, and can be beneficial to them insofar as the test's answer key is middle biased. In any case, because test takers' middle bias is not subject to control by test makers, we argued that it needs to be acknowledged by them when estimating psychometric parameters of items for operational use.

Although "response sets always lower the logical validity of a test. Empirical validity, based on ability to predict a criterion, may be raised or lowered" (p. 484), yet "Even though the empirical effect may be small, [Cronbach, 1946] feels that response sets should be eliminated where possible" (p. 487). The testing establishment has almost uniformly adopted key balancing as the manner to eliminate position biases. We have critiqued this solution, showing its inferiority to key randomization—but that is a matter for a separate article (Bar-Hillel & Attali, 2001).

Notes

[1] We thank Dr. Danny Cohen for directing us to this quotation.

[2] These constitute all the 25-question sections among the 8,905 PET questions piloted in 1997 and 1998 (see Table 2).

[3] NITEST (Cohen & Bodner, 1989), the program that is used at NITE for item parameter estimation and calibration, is modeled after that of ASCAL (Vale & Gialluca, 1985).

[4] The four-choice SAT questions with 47% correct middle answers are also, of course, composed of questions that underwent pretesting, but because we have no access to the ETS' pilot bank, we can only tentatively offer the same interpretation.

References

Ayton, P., & Falk, R. (1995). *Subjective randomness in hide-and-seek games.* Paper presented at the 15th bi-annual conference on Subjective Probability, Utility, and Decision Making, Jerusalem, Israel.

Bar-Hillel, M., & Attali, Y. (2001). *Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests.* Manuscript submitted for publication.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics, 12,* 428–454.

Bereby-Meyer, Y., Meyer, J., & Budescu, D. (2000). *Decision making under uncertainty and partial knowledge: The case of multiple-choice tests.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Berg, I. A., & Rapaport, G. M. (1954). Response bias in an unstructured questionnaire. *The Journal of Psychology, 38,* 475–481.

Christenfeld, N. (1995). Choices from identical options. *Psychological Science, 6,* 50–55.

Claman, C. (1997). *10 real SATs.* New York: College Entrance Examination Board.

Clark, E. C. (1956). General response patterns to five-choice items. *Journal of Educational Psychology, 47,* 110–117.

Cohen, Y., & Bodner, G. (1989). *A manual for NITEST—A program for estimating IRT parameters* (CAT Project Report No. 1). Jerusalem, Israel: National Institute for Testing and Evaluation.

Cronbach, L. J. (1946) Response sets and test validity. *Educational and Psychological Measurement, 6,* 475–494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10,* 3–31.

Fagley, N. S. (1987). Positional response bias in multiple-choice tests of learning: Its relation to testwiseness and guessing strategy. *Journal of Educational Psychology, 79,* 95–97.

Falk, R. (1975). *The perception of randomness.* Unpublished doctoral dissertation (in Hebrew, with English abstract), Hebrew University, Jerusalem, Israel.

Gibb, B. G. (1964). Test-wiseness as secondary cue response (Doctoral dissertation, Stanford University). Ann Arbor, MI: University Microfilms (No. 64-7643).

Graduate Management Admission Council. (1992). *GMAT review, The official guide.* Princeton, NJ: Author.

Gustav, A. (1963). Response set in objective achievement tests. *Journal of Psychology, 56,* 421–427.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2,* 51–78.

Inbar, N. (2002). *Middle bias in numerical answers to multiple-choice questions.* Unpublished master's thesis, The Hebrew University, Jerusalem, Israel.

Instituto Nacional de Estudos e Pesquisas Educacionais. (1998). *Avaliacao de concluintes do ensino medio em nove estados, 1997: Relatorio final.* O Instituto.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Keillor, G. (1997). *Wobegon boy.* New York: Penguin.

Kiddum. (1995). *Preparation for the Psychometric Entrance Test.* Israel: Author.

Kubovy, M., & Psotka, J. (1976). Predominance of seven and the apparent spontaneity of numerical choices. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 291–294.

Marcus, A. (1963). The effect of correct response location on the difficulty level of multiple-choice questions. *Journal of Applied Psychology, 47,* 48–51.

McNamara, W. J., & Weitzman, E. (1945). The effect of choice placement on the difficulty of multiple-choice questions. *Journal of Educational Psychology, 36,* 103–113.

Mentzer, T. L. (1982). Response biases in multiple-choice test item files. *Educational and Psychological Measurement, 42,* 437–448.

Metfessel, N. S., & Sax, G. (1958). Systematic biases in the keying of correct responses on certain standardized tests. *Educational and Psychological Measurement, 18,* 787–790.

Offir, R., & Dinari, N. (1998). *Chemistry for matriculation exams.* Tel Aviv, Israel: Lachman.

Open University, The. (1998). *Psychological development: A study guide.* Tel Aviv, Israel: Author.

Rapaport, G. M., & Berg, I. A. (1955). Response sets in a multiple-choice test. *Educational and Psychological Measurement, 15,* 58–62.

Rubinstein, A., Tversky, A., & Heller, D. (1996). *Naive strategies in competitive games.* In W. Albers, W. Guth, P. Hammerstein, B. Moldovanu, & E. van Damme (Eds.), *Understanding strategic interaction.* New York: Springer-Verlag.

Shaw, J. I., Bergen, J. E., Brown, C. A., & Gallagher, M. E. (2000). Centrality preferences in choices among similar options. *The Journal of General Psychology, 127,* 157–164.

Trivia. (1999). Retrieved from http://www.triviawars.com

Ullmann-Margalit, E., & Morgenbesser, S. (1977). Picking and choosing. *Social Research, 44,* 757–785.

Vale, C. D., & Gialluca, K. A. (1985). ASCAL: *A microcomputer program for estimating logistic IRT item parameters* (ONR-85-4). St. Paul, MN: Assessment Systems Corporation.

Yoel, A. (1999). A collection of questions from the matriculation tests in biology. Tel Aviv, Israel: Lachman.

Authors

YIGAL ATTALI, Educational Testing Service, Rosedale Road MS-10-R, Princeton, NJ 08541; yattali@ets.org. His research interests include educational measurement and cognitive aspects of testing.

MAYA BAR-HILLEL is a professor, Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, Israel 91905; maya@huji.ac.il. She is also the director of the Hebrew University's Center for the Study of Rationality. Her research interests include rationality, probabilistic reasoning, and judgment.