דוח המרכז

286

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

Yigal Attali
Maya Bar-Hillel

June 2001

מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
מיסודן של האוניברסיטאות בישראל
ת"ד 26015 ירושלים 91260, טל' 02-6759555

דוח מרכז 286

ISBN:965-502-063-0

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

Yigal Attali

National Institute for Testing and Evaluation

The Hebrew University of Jerusalem

&

Maya Bar-Hillel

The Hebrew University of Jerusalem

Paper submitted to the Journal of Educational Measurement

June 2001

Acknowledgements

We wish to thank the National Institute for Testing and Evaluation for its continuing support and encouragement, without which this work could not have been carried out.

We would like to thank those who read and commented on earlier versions of this manuscript, and assisted the first author as a graduate student, including: Michal Beller, Gershon Ben-Shakhar, David Budescu, Danny Cohen, Yoav Cohen, Naomi Gafni and Shmuel Zamir.

Abstract

In this paper, we 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 over 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 to a lesser extent than the correct key. In a typical 4-choice test, about 55% of erroneous answers (which are the only answers whose position is determined by the test taker, not the test maker) are in the two central positions. We show that this bias is large enough to have real psychometric consequences, as questions with middle correct answers are easier and -- what's more important -- less discriminating than questions with extreme correct answers, a fact some of whose implications we explore.

“ *[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"* ”

From: Garrison Keillor (1997) Wobegon Boy. Penguin Books (p. 180) [1].

This paper, and its companion paper (Bar-Hillel & Attali, 2001), explore the role of 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 present paper discusses within-item position effects. The companion paper deals with sequential (across-items) position effects.

I. Where do people place the correct answer when constructing a single multiple -choice question? Apriori 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? Apriori, 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. Since the question and its correct answer seem to be the 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 the 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 than itself (Inbar, 2001). Since 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 towards the center as a kind of compromise, or "center of gravity", of the set of possible positions. The center may also feel like a less conspicuous place to "hide" the correct answer (see some evidence for this in Section II). If so, there would be a preponderance of correct answers in central positions. This paper 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 -- an absence of any 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

As to individual randomization, it is well known that people are quite incapable of simulating a random device in their heads, and the only way they can randomize is 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, and the topic deserves some elaboration.

Consider a multiple-choice test as a zero-sum game of "hide-and-seek" (see Section IV. below, and Rubinstein, Tversky & Heller, 1996). The examiner wants to hide the correct answer, especially from a guessing examinee, and the examinee wants to find it, even when guessing. An elementary result from Game Theory is that this game has a unique Nash equilibrium with mixed strategies: the position of the correct answer should be picked at random (i.e., by a lottery giving equal probabilities to each position). Any other strategy can be exploited by a test taker who fathoms it.

The intuition behind the equilibrium theorem of Game Theory is readily explained. The examiner wants to "hide" the correct answer where the ignorant examinee will not find it -- or will be least likely to find it. The examinee wants to find it, ignorance notwithstanding, and so tries to figure out where an examiner, who does not want him to find it, is likely to hide it. Any position that is a "good" place to hide the correct answer is, perforce, also "bad", since what makes the examiner regard it as "good" can be figured out by the examinee, which is rather "bad". That said, the examiner will avoid it -- but then, so will the examinee. And so on, ad infinitum. The only way out of this quandary is for the examiner to randomize the position of the correct answer -- a strategy that cannot be exploited by the examinee even if it is public knowledge. It guarantees the examiner that the examinee can have, on positional grounds alone, no better than a 1/k probability of getting the right answer, while simultaneously guaranteeing the examinee at least a 1/k probability of that very possibility.

Although Game Theory treats "hiders" and "seekers" symmetrically, there is a basic psychological asymmetry between them, deriving from the fact that the former "move" first. Test makers are free to place the correct answer wherever they choose. Test takers, on the other hand, want to mark the correct answer where it actually is. If test takers can figure out where the correct answer is, by whichever means, they mark it. Usually, they figure out *where* it is simply by figuring out, or recognizing, *what* it is. A testwise examinee can sometimes figure out where the correct answer is even without knowing (test-wiseness is the "capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score", Millman, Bishop & Ebel, 1965, p. 707), and test makers are admonished to be aware of this possibility and to try to minimize it. The way to minimize the possibility that answer position could be exploited by a testwise examinee is to randomize it.

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.

II. Where do people place the correct answer when constructing a single multiple -choice question? Empirical evidence

This question cannot be answered merely by looking at published tests, because when constructing an entire test, test makers often move the correct answer from the position in which it appeared originally, in order 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 "[N]o directions concerning answer placement were given" (p. 477). In 1000 consecutive cards from the first assignments, correct answers were distributed over the four positions as follows: [54, 237, 571, 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 which had numerical answers. Answer positions were labeled A, B, C, and D -- but left empty. The respondents were a convenience sample consisting all 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, 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 of the authors, were either students and faculty in The Hebrew University's Psychology Department who were approached in departmental seminars, or personal friends who were approached in dinner parties, and 7 English speaking academic colleagues of the second author who were approached by e-mail. Except for the latter, the request was delivered orally and informally, and questions were jotted on the respondents' own pieces of paper. The academic groups, most of whom had extensive experience in writing multiple-choice tests for their students, were instructed to write a 4-choice question on any topic they wished, avoiding answers of the sort: "All answers are true" or "No answers are true". The personal friends were told to "Write a question for [the popular TV program] 'Who Wants to be a Millionaire' ", and apparently most believed their questions might actually be passed on to the program's producers. None of the people had any notion what this was about, and no explanation was offered. Since they did not differ on the dependent variable, we report on all together. The distribution of correct answers over positions was [7, 21, 30, 7] -- nearly 80% in central positions ($p<.0001$).

We reported separately on the 125 strangers and the 65 acquaintances, because we debriefed many of the latter, but none of the former, thus obtaining some informal observations. None suspected that the point of the little exercise was to see where they put the answer, and they were quite surprised to hear its true purpose -- and doubly surprised at the results. Their guesses were that we were looking to see what topics people choose to ask about, etc. They admitted to 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 they seemed to have positioned the correct answers with little if any deliberation.

III. 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 in whichever position 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 is a test taker likely to consider position. The need to identify guesses makes it harder to determine where people "seek" the correct answer to a multiple-choice question then to determine where they "hide" it.

To ensure guessing, a class of 127 undergraduate Psychology students were 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. 69 respondents 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, 1]. In the subsequent question, it was [11, 21, 26, 11] if the first question had been self-answered (N=69), and [12, 20, 22, 4] if the first question had been pre-answered (N=58). In toto, the percent of times each answer position was chosen over the 196 choices (69 + 69 + 58) made in both conditions and both questions was [15%, 38%, 39%, 8%] -- hence almost 80% middle choices. Only 50% would have been expected by chance (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 9 various kinds of forced-choice "imaginary questions". In their "question" II the answers were labeled 1, 2, 3, 4, and in question VIII, they were labeled A, B, C, D. In an inverted form given to 203 other students, the labels were 4, 3, 2, 1 and D, C, B, 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. A group of 40 Technology students were asked "What does the Japanese word TOGUCHI mean?". None of these Israeli respondents knew any Japanese. The possible answers, and their response frequencies were: A. Door (8) B. Window (23) C. Wall (3) D. Floor (6) -- 65% middle answers. When asked to explain why they had placed the correct answer where they did, only two people gave a reason that mentioned position explicitly ("I just gave the first answer"; and "C just grabbed me").

The final piece of evidence we collected regarding how people guess in a single question comes from analyzing responses in a particular kind of guess, given within a real test. The data for this analysis, as well as for many of those to follow, derive from the Psychometric Entrance Test (PET), a 4-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, like the SAT, 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 adults (over 90% are under 26) with a high-school education (those who have not graduated yet will have to before they can be admitted to the universities).

In the PET's quantitative sections, questions are ordered from easy to difficult. For that reason, as well as for the reason that the tests are timed and respondents may run out of time towards the later questions, some test takers appear to "give up" towards 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 it exclusively from some point in the test on. It could be argued that perhaps some of these runs are not the consequence of "giving up" but simply reflect the test taker's best guess, but the longer the sequence of identical final answers, the less likely is that to be the case. Admittedly, long runs of identical answers sometimes (but again, very rarely) appear in the middle of sections, and not only towards the end, and these may be harder to interpret as just "giving up", since they are abandoned before the end.

Be that as it may, 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 a total of 71,120 25-long answer sequences. Table 1 shows tail-runs (i.e., sequences of identical responses ending in the final item), other runs (i.e., sequences of identical responses ending before the final item), and the percent thereof in which a central position was chosen. The first row shows this percent among all runs of length 4 (that being the shortest run that never appears in the answer keys; see Bar-Hillel & Attali, 2001), the second shows it among runs of length 5, etc. 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 some run is really a "give up" guess (or: the higher the percent 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 represent a single choice -- the one starting the run.

Table 1

Percent of middle choices in long runs

Run length	N of tail runs	% in middle	N of other runs	% in middle
4	1266	64	7118	67
5	534	78	1664	74
6	184	83	440	75
7	97	91	161	84
8	49	84	69	86
9	24	83	33	91
≥10	30	87	46	81

All in all, we have shown that people writing isolated 4-choice questions hide the correct answer in the two middle positions about 70% of the time, and people guessing an isolated 4-choice question seek it in a middle position about 75% - 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 quote), 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). Indeed, we only found tips recommending guessing a central position in some Internet sites which give testwise cues for dealing with ordinary multiple-choice tests, not the professional high stakes ones. We ourselves were quite surprised by the magnitude of the edge aversion we found, and so were most of our question writers, once briefed. 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 (see section V below). 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 75% of the time, as we reported for single questions -- and certainly not just 55% of the time, as we shall report below (Section VI) for guessing in real tests. On the other hand, the probability-matching response (e.g., Estes, 1976) is to guess middle answers about 52% of the time (the typical middle bias in many tests, see Table 2), which is a bit less than the 55% rate we found (see Section VI.). Last but not least, in the following section we show that a tendency towards 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.

IV. Edge Aversion

In a 4-choice test such as the PET, it is hard to say whether the marked preference for placing correct answers in the middle should be attributed to an attraction to the middle, or to an aversion to the edges. But a similar bias has been observed in many other tasks, which we survey below. In some it is clearly edge aversion rather than (strict) middle attraction which causes the bias (2,3 and 5 below).

1. The closest task to a multiple-choice test was studied by Rubinstein, Tversky and Heller (1996). They instructed subjects to "hide a treasure" in one of four places laid out in a row, where other subjects would then look for it by getting a single opportunity to observe the content of a single hiding place. They were told that the "hider" wins if the treasure is not found, whereas the seeker wins if it is. Both "hiders" and "seekers" favored middle positions. The analogy of the latter task with a multiple-choice test is obvious. The authors seem to have shared our intuition regarding what drives this bias, talking about "players' tendency to avoid the endpoints" (Rubinstein, Tversky & Heller, 1996, p. 399). The fact that the bias was common to "hiders" and "seekers" undermines any notion that it is somehow related to strategic advantage.

2. Falk (1975) asked subjects to mark 10 cells in a 10 x 10 matrice "as if these cells were chosen by blind lottery". The cells on the border of the grid were heavily avoided. The median (and modal) number of edges which the border shared with the marked cells was 2 -- half the number expected from marking cells at random. However, within the 8 x 8 interior array, the middle was not systematically preferred, lending further support to the notion that the bias towards the middle is really a bias away from the edges.

3. In a kind of extension of both Falk's earlier work and that of Rubinstein, Tversky and Heller's, Ayton and Falk (1995) asked respondents to mark 3 cells in a 5 x 5 matrice 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 (16 to 9 with no cell exclusions, 12 to 8 with the corners and the middle excluded).

4. 5-choice tests such as the SAT allow a distinction between middle-attraction and edge aversion. In 5-choice tests, while positions A and E are the least popular, position C -- the

precise middle -- is not the most popular, suggesting that it is not so much attraction to the middle as aversion to the extremes which underlies middle-bias (see Table 2 below).

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. Our physical world being what it is, otherwise identical objects occupy different places in space-time. Coffee jars, for example, necessarily occupy different places on the supermarket shelves. In a study called "Choices from identical options", Christenfeld (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). Middle toilet stalls were chosen 60% of the time (as estimated from the amount of toilet paper used).

The latter study is particularly interesting, since the real life contexts it studied had different strategic structures. Though the tendency of one's fellowfolk to select products from the middle of a grocery shelf makes it strategically advantageous for oneself to do the same (since products there will tend to be replaced more often, hence be more fresh), whereas the tendency of one's fellowfolk to use central bathroom stalls makes it strategically disadvantageous for oneself to do the same (since 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 (they would have been chosen 10% of the time by random choice). The favorite choices were 7 (nearly 30%), followed by 3 (nearly 15%). These results indicate edge aversion rather than middle-favoring -- indeed, 4, 5 and 6 were not as popular as 7 and 3, and only about as popular as chance.

Although in the opening section we derived the prediction of a middle bias under a heading of "compromise", we have no direct evidence for anything of the kind, and the middle bias in tests is most probably nothing more than a manifestation of the general phenomenon of edge aversion. There is no agreed upon explanation for edge aversion (or bias towards the middle) in its general form, and in spite of some attempts to explain it (e.g., Shaw, Bergen, Brown & Gallagher, 2000), it is a phenomenon still in search of an explanation. We will use the terms middle bias and edge aversion interchangeably, in a manner uncommitted to any underlying causal mechanism.

V. 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 4-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? Do lengthy sequences still exhibit edge aversion? This section will address this issue from the perspective of test makers, and the following section will do so from the perspective of test takers.

The dominant, if not publicly acknowledged, policy regarding answer keys, favored by professional test makers, is known as key balancing (see Bar-Hillel & Attali, 2001). Notably, this is the policy at one of the world's largest testing organizations, the ETS (the Educational Testing Service, makers of the SAT and the GRE, among others). There is global key balancing, and local key balancing. Global balancing involves i. making sure that the various positions appear in roughly equal proportions in the answer key of each subset. Local balancing involves ii. making sure that the correct answer never appears in the same position more than three times in a row. In addition, iii. Correct answers must appear once in every position, even in relatively small "windows" (i.e., sequences of consecutive questions) -- about 9-10 long in 4-choice tests such as the PET, about 15 long in 5-choice tests such as the SAT. See more detail in Bar-Hillel and Attali (2001).

These policies notwithstanding, the psychometric literature has little to say about key balancing (perhaps because intuitions about its benefits and practice are so widely shared).

Indeed, it has all but ignored the positioning of answer options within a question as a characteristic of interest, much less as one that could affect the psychometric indices of either individual items or entire tests. No psychometric model formally incorporates answer position, seeming to take it for granted that position cannot affect test item characteristics.

As a result of global balancing, any edge aversion which 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, and we regard them as representative, at least in regards to bias in the answer key. NITE gave us access to answer keys of some tests they developed (e.g., 8905 PET questions piloted in 1997-1998; 10 operational PET exams; MMI -- a Mathematics exam constructed by NITE for student selection at Israel's Technion in 1988-1999). Correct answers to trivia questions were taken from the Internet site www.triviawars.com. in the fall of 1999. The other answer keys appeared in published sources, as cited (e.g., SAT answer keys from Claman, 1997). Of the 13 sources we considered, 11 produced tests with an over-preponderance of middle answers (p=.01, sign test).

Table 2

Percentage of correct answers by positions in various answer keys

Test	# of Questions	A	B	C	D	E	% in middle
4-choice tests							
PET pilot [a]	8905	25	26	25	24	-	51 [b]
10 Operational PET tests [a]	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 U (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 tests							
SAT (Claman, 1997)	1130	19	20	22	21	19	63 [b]
MMI[a] (1988-1999)	1440	18	22	21	21	18	64 [b]
INEPE (1998)	432	18	25	21	19	18	64 [b,c]
GMAT (1992)	402	17	19	23	22	19	64

a Answer keys courtesy of the NITE
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.

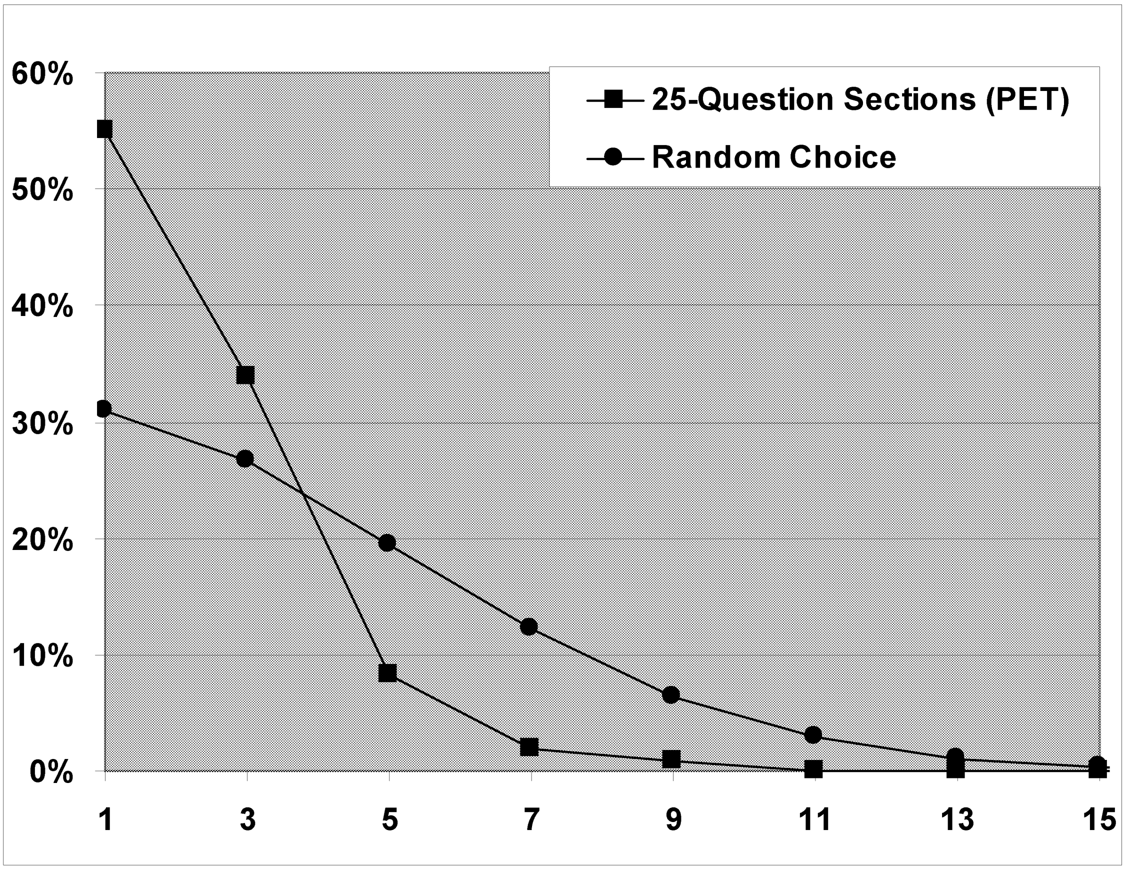
We found three previous surveys of a similar kind. Metfessel and Sax's (1958) results are hard to quantify, but they summarized them in the words: "There is a tendency for the authors of multiple-choice tests to place the correct answer at the center" (p. 788). McNamara and Weitzman (1945) analyzed nearly 5000 5-choice questions and 4000 4-choice questions taken from operational tests, and found correct answers placed in center positions in 62% of the former and 51% of the latter. Finally, Mentzer (1982) presents a table much like ours, reporting the distribution of correct answers over positions in a 4-choice test, and in 30 of 35 tests, the answer was in central position (p<.0002, sign test).

An "anchor and adjust" procedure could explain these findings. When correcting for the preponderance of correct answers in middle positions, answer positions are permuted, moving correct answers from the middle to the edges. This iterative process can be stopped when the proportions are roughly -- and not necessarily precisely -- equal. But starting, as it does, with a preponderance in the center, chances are that the process will halt when this preponderance, though considerably diminished, is still there.

.

Figure 1

Distribution of absolute difference between middle and edge positions in 25 -questions PET sections



Ironically, while the mean difference between the observed percent of middle responses and the expected percent is positive, hence too large, yet 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 4-choice tests; 64% in the 5-choice tests), the variability around the proportion is too small: there are not enough sections in which there is a major imbalance between, say, middle positions and edge positions. For example, Figure 1 shows the distribution of sections with a given (absolute) difference between middle versus edge positions in 109 4-choice PET sections [2]. Since the sections analyzed are all 25 items long, this number ranges between 1 (13 versus 12) to 25 (25 versus nil). The figure shows the distribution expected under random positioning of the correct answer, and the observed distribution under a policy of key balancing. Small differences (1 and 3, i.e., the difference 13-12 and 14-11) are over-represented in the actual tests, and large differences (5 and up) are under-represented

Because the SAT sections are of different lengths, we could not produce a similar figure for the SAT. So we measured over-balancing in the SAT key as follows. For each of the six sections that compose an SAT test, we computed the absolute difference between the expected number of extreme positions in a section (assuming random choice) and the observed number. For example, in a section of 25 questions, if 8 correct answers are in position A or E, rather than the expected 10, then the absolute difference is 2. These absolute differences were summed over all six sections of 10 real SAT tests (Claman, 1997), and are presented in the first row of Table 3. We next conducted a Monte Carlo simulation on 100,000 randomly generated SAT answer keys, and computed the sum of the absolute deviation (SAD) for each. This yielded a distribution of SADs whose mean was 10.5 (SD=3.4), and whose median was 10.3. The second row of Table 3 shows the percentile rank of the SADs of the 10 real SAT keys in the expected distribution of SADs. One can see that all but one of the 10 observed SADs has a percentile rank of less than 50 (p=.01, sign test), and the median percentile rank is 15.

Table 3

Observed SADs in 10 real SAT test keys, and their percentile rank in the expected distribution under randomization SAD

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

Merely for comparison ease, we carried out the same SAD analysis for the PET, using the 10 operational PET tests mentioned above. The median SAD 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 (namely, too close to uniformity) and not balanced enough (namely, still middle-biased). The former is the deliberate and intended consequence of global balancing, and would be expected wherever 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.

VI. Are examinees' in real tests middle biased?

Since test makers produce balanced keys, test takers, insofar as they mimic them, should also produce balanced keys. High ability test takers are, by definition, those who were successful in reproducing the test maker's answer key with little or no error. The higher one's ability, the less one errs. Hence, a test taker has an opportunity to exhibit edge aversion only with respect to items on which she has imperfect knowledge, if any.

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 which 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 (quite rightly, as it happens) that the correct answer key is balanced, may wish to produce a representative answer sequence on their own, even if they cannot quite reproduce the correct one. In other words, they may believe (and again, quite rightly so) that they increase their chances of answering correctly if they guess in a representative manner, namely, produce an answer sequence whose properties reproduce those of the correct answer key.

These two robust, and documented, tendencies -- aversion to edge positions, and representativeness -- conflict when guessing in a sequence of questions, since the first leads to unbalanced guessing (middle bias), and the latter to balanced guessing. We carried out a series of analyses on real test results, as well as on experimental data, to see whether, and to what extent, test takers 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 wouldn't be any "response set", in his terminology, and others have concurred in this expectation. But when considered empirically, "The 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 ourselves looked at the test results of the examinees who took the 8905 questions piloted by the NITE in 1997-1998 (about 300 answers per question). Table 4 shows the percent of all middle answers, 53%, as well as the percent of just erroneous middle answers (errors constituted just over a third of the total), 55%. The test takers, we see, produced answer sequences with a middle bias only slightly higher than observed in the correct answer key, 51% (from Table 2).

Table 4

Percent of examinee answers in each position.

	A	B	C	D	middle
All answers	23	26	27	23	53
Erroneous	22	27	28	23	55
Correct key	25	26	25	24	51

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 is no correct answer (e.g., "Where does the Lagundo tribe live?: a. West Africa b. New Guinea c. Ecuador d. Indonesia" -- where there is no Lagundo tribe ....). The 144 respondents preffered middle options to extreme ones 53% of the time (Bereby-Meyer, Meyer & Budescu, 2000).

Switching the position of answer options

If the attraction of a guessing examinee to different answer options is influenced by their position, changing the positions of the answer options should reveal this bias. Specifically, suppose one group of examinees receives questions with the answer options in positions A, B, C, D, and another group gets the same questions with the same answers, but reordered into positions B, A, D, 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 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 4000. The examinees who participated in this experiment did so in the course of taking the PET exam, which hides a pilot section among the operational sections used for university selection.

Two dependent variables were used -- the difference between the percentage of examinees choosing an option when it was presented in a middle position and when it was presented in an extreme position, and the ratio of these percentages. This analysis focuses on wrong answers, on the general assumption that these are mostly guessed answers. On average, the wrong options attracted about one third of the examinees, so about 11% of them chose each one. Table 5 shows that the mean difference in the percent 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.

Table 5

Difference between, and ratio of, the percent of examinee s that 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

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 who 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 will 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). Note that *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 position, 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,

I. $\mathrm{P(C}|E\mathrm{)} = \mathrm{P(K)} \cdot 1 + \mathrm{P(G)} \cdot \mathrm{P(E|G)} / 2 = (1 - \mathrm{P(G)}) + \mathrm{P(G)} \cdot (1 - \mathrm{P(M|G)}) / 2$

II. $\mathrm{P(C}|M\mathrm{)} = \mathrm{P(K)} \cdot 1 + \mathrm{P(G)} \cdot \mathrm{P(M|G)} / 2 = (1 - \mathrm{P(G)}) + \mathrm{P(G)} \cdot \mathrm{P(M|G)} / 2$

Since 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, which we assume 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 4-choice question equals 1/2, whereas our research hypothesis is that it is larger than 1/2.

Isolating P(M|G) from I. and II. we obtain:

III. $\mathrm{P(M|G)} = (1 + \mathrm{P(C}|M\mathrm{)} - 2\mathrm{P(C}|E\mathrm{)}) / (2 - \mathrm{P(C}|M\mathrm{)} - \mathrm{P(C}|E\mathrm{)})$.

For example, when P(C|*M*) is .67, and P(C|*E*) is .64, then P(M|G) is 0.56. Indeed, these were the mean values for P(C|*M*) and P(C|*E*), respectively. But we used III. to estimate P(M|G) separately for each of the 161 questions in the switching experiment. The resulting estimate was a significant 0.57 ($t_{160} = 6.9$, $p < .0001$). P(M|G) as estimated from this model is an under-estimation, because the model simplistically ignores examinees whose errors are not due to guessing (e.g., because they are due 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 the same 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 III. 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 (more on this can be found in Bar-Hillel & Attali, 2001).

VII. Psychometric consequences

Edge aversion is a bias with which psychometrics must contend, as it has implications for psychometric indices such as item difficulty and item discrimination. Since 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.

Position effects on item difficulty and discriminability: Evidence from the switching experiment

Table 6 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.

Table 6

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

	N of items	Mean	SD	t value
Difficulty (drop in % correct)	161	3.3%	5.1	8.1*
Discriminability (increase in the Biserial)	161	0.05	0.1	5.5*

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

Note that the effect of position (middle versus extreme) on correct responses, 3.3%, is even larger than its effect on incorrect responses (see Table 5, 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 number of guessers. Correspondingly, the position effect can be expected to be larger, too. Indeed, the regression models for predicting both indices -- difficulty (measured by percent correct) and discriminability (measured by the biserial correlation) -- from the mean difficulty of the item (over the two presentations) are significant (p<.01). Table 7 presents the predicted values of the position effects.

Table 7

Predicted effects of moving the correct answer from a middle to an extreme position on item indices of 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

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 8905 PET questions that were 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). Note that 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 real tests conventionally place them in ascending order of difficulty. Table 8 presents the mean effect for the Quantitative questions by the position of the question in the section -- a proxy for their difficulty.

Table 8

Percent of correct answers in middle and in extreme positions, by question's ordinal position in test

	Ordinal position of questions within the section				
	1-6	7-12	13-18	19-25	All
Middle	72	65	60	50	62
Extreme	71	62	54	43	57
Difference	1	3	6	7	5

The gain in percent correct due to the middle position bias rises from 1% to 7% as difficulty increases -- a wider range than predicted by the regression model (Table 7 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 whose biserial value for the correct answer is 0.3 or higher, and whose point-biserial values for all distractors is -0.05 or lower. It seems odd that the quality of a question would depend on the position of the correct answer. Yet Table 9, in which the questions were sorted into quartiles by difficulty, shows the difference in percent of "good" questions among those with middle versus extreme correct answers. For both types, it presents the percent of successful questions, the percent 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.

Table 9

Percent of "good" questions, and mean biserial, by position of correct answer and difficulty quartile

	Middle correct position					Extreme correct position				
	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

Position effects on IRT based item indices: Evidence from operational PET questions

A final analysis examined position effects on IRT based item indices. At NITE, IRT parameters of PET items are estimated only for operational items, because piloted items are not administered to enough examinees (typically 200-300 only). As a result, we conducted this analysis on operational PET items that are administered typically to several thousand examinees [3]. Since operational items are selected on the basis of their discrimination, we expect that any position effects will be smaller in operational items than in piloted items.

We compared item parameters of all operational PET items from the years 1997-1999, which amounted to 4533 questions. The next table shows means of the a, b and c parameters for middle-keyed and extreme-keyed questions. For all three parameters a, b and c 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 than 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 tenth of the standard deviation for a, to about half of the standard deviation for c.

Table 10

Means (and SD) of IRT item parameters (based on PET 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		.09	.17	.49

Position effects on successful guessing

One component of question difficulty is the probability of guessing it correctly. The shared preference of examinees and examiners for placing correct answers in middle positions brings about a positive correlation between the choices of guessing test takers and the choices of test makers. This positive correlation is, of course, advantageous to the examinee, because it enhances the probability of a successful guess. If the distribution of percent of guessing examinees over positions is (P_1, P_2, P_3, P_4) and the distribution of the percent of examiners positioning of correct answers is (Q_1, Q_2, Q_3, Q_4), then the probability of a match between the choices -- a match that leads 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 bias toward middle options is a typical 52% in the key, and 55% for examinees, then the match probability in a 4-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%. In classroom tests written by lay people who may fail to balance the key, the middle bias may be intermediate between perfect balancing and the one-item bias, and the match probability will be correspondingly intermediate between 25% and 40%.

The beauty of randomization is that if the test makers randomize, thereby producing a bias-free answer key (on average), it doesn't matter whether the test takers are biased or not -- their probability of a match will be 25%. In that respect, if answer keys were randomized, it wouldn't 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 untested 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 9, 51% of the edge-keyed Q4 questions, but only 34% of the middle-keyed Q4 questions, are "good". Hence, if one starts, for example, with 100 edge-keyed difficult items and 100 middle-keyed difficult items, one will end up with 85 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.

VIII. Conclusion

In the present paper we argue that the "hiding" of correct answers in the possible positions should be done at random. The case for randomizing the entire answer key is made in our companion paper (Bar-Hillel & Attali, 2001), but is, in principle, the same. Both rest on the game-theoretic notion of equilibrium. A defining property of the equilibrium in general, and in the case of multiple-choice tests in particular, is that it cannot be exploited, even if it is common knowledge. A side benefit of a key-randomizing policy is, therefore, transparency: it can -- and should -- be publicly acknowledged. The ETS's (and ACT's [5]) present policy on answer keys is a "trade secret", in the sense that it is not publicly acknowledged. But key randomization on the part of test makers does not do away with the middle bias of test takers. Since that bias cannot be controlled by test makers, it needs to be acknowledged by them.

Notes

1. We thank Dr. Danny Cohen for directing us to this quote.

2. These constitute all the sections of length 25 among the 8905 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 4-choice SAT questions with 47% correct middle answers is also, of course, composed of questions that underwent pretesting, but since we have no access to the ETS's pilot bank, we only tentatively offer the same interpretation.

5. In April, 2000, following a presentation of this paper at the annual meeting of NCME in New Orleans, the ACT discussant commented wryly that when he asked at ACT whether key balancing was being practiced there, they refused to answer.

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, August 1995.

Bar-Hillel, M., & Attali, Y. (2001). Seek Whence: Answer sequences and their consequences in key-balanced multiple-choice tests. *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. Submitted for publication (paper presented at the 2000 NCME meeting (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. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*, 3-31.

Estes, W.K. (1976). The cognitive side of probability learning. *Psychological Review, 83*, 37-64.

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, Michigan: University Microfilms, No. 64-7643.

GMAT (1992). *GMAT Review, The Official Guide*. Princeton, NJ: Graduate Management Admission Council.

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. (2001). Middle bias in numerical answers to multiple-choice questions. MA thesis, The Hebrew University, Jerusalem.

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.

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

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.

Millman, J., Bishop, C.H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707-726.

NITE (1998). *Psychometrics any way you test it (in Hebrew).* Jerusalem: The NITE.

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

Open University, The (1998). *Psychological development: A study guide (in Hebrew).* Israel: The Open University.

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). 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: Assessment Systems Corporation.

Yoel, A. (1999). *A collection of questions from the matriculation tests in Biology (in Hebrew).* Israel: Lachman.