האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

# SEEK WHENCE: ANSWER SEQUENCES AND THEIR KEY-BALANCED MULTIPLE- IN CONSEQUENCES CHOICE TESTS

by

MAYA BAR-HILLEL and YIGAL ATTALI

Discussion Paper # 252

Following the 2002 publication you can find the fuller 2001 original manuscript bearing the same title.

מרכז לחקר הרציונליות

CENTER FOR THE STUDY
OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

# Seek Whence: Answer Sequences and Their Consequences in Key-Balanced Multiple-Choice Tests

Maya Bar-Hillel and Yigal Attali

The producers of the SAT balance answer keys rather than randomizing them. Whereas randomization yields keys that are balanced only on average, balancing assures this in every subtest. Balancing is a well-kept trade secret, and there is no evidence of awareness that it is exploitable. However, balancing leaves identifiable traces on answer keys. We present the evidence for key balancing, its signatures, and the ways in which testwise examinees can exploit it. Exploitation can add as much as 16 points to one's SAT score.

KEY WORDS: Randomization; SAT; Testwiseness.

## 1. "THE DELICATE ART OF KEY BALANCING," OR: WHEN RANDOMIZATION IS TOO IMPORTANT TO BE TRUSTED TO CHANCE.

Surprisingly, people writing a multiple-choice question tend to place the correct answer in a central position up to three to four times as often as at an extreme position, with little if any awareness of this tendency (Attali and Bar-Hillel in press). Banks of multiple-choice questions therefore usually exhibit a preponderance of answers in middle positions. If the correct answers are not reassigned to different positions, the resulting answer key could be heavily unbalanced. The near-universal method of dealing with this bias is through the so-called "delicate art of key balancing." Key balancing is not an openly practiced policy (when a reader of this article asked whether key balancing was practiced at the ACT, where he works, they refused to answer), and its secrecy is maintained for good reasons—the same reasons that should have militated, as we shall see, against the very practice.

Key balancing was, until recently, the unwritten answer key policy at NITE, Israel's National Institute of Testing and Evaluation. NITE produces and administers the Psychometric Entrance Test (PET), which measures various scholastic abilities and achievements, and is used for student admissions by all Israeli universities. It resembles the SAT, developed for similar purposes by the Educational Testing Service (ETS), but it is a four-choice test (the SAT is mostly five-choice) consisting of two sections of size 25 (Quantitative), two of size 27 (Verbal), and two of size 30 (English). In 1999, as a result of the present work, NITE abandoned key balancing in favor of key randomization. The following rules of thumb characterized its (then-secret) key policy regarding sections of 25 questions:

1. No position should appear in the section key more often than nine times, or less often than four.

2. Correct answers should never be placed over three times in a row in the same position.

3. A sequence of about a half the length of the section (i.e., about a dozen consecutive items) should not lack one of the four positions.

Rule 1 can be called "global balancing," albeit at a section level, while Rules 2 and 3 are more local balancing practices. Global balancing, Rule 1, once achieved, is independent of item reordering, whereas run avoidance, Rule 2, and position-neglect, Rule 3, the local properties, depend on the actual sequencing of the questions. All of the following answer keys, though seemingly "random," would be ruled out:

A B C B A B B D C B A B D B A C C C A B C D C A A

(Only three D's, violates Rule 1.)

A B C D A B B D C B A B D B C C C C A B D C C A A

(Run of four C's, violates Rule 2.)

A B B C B A A A C B C C C D D C A A D D D A B B D

(No D in first half, violates Rule 3.)

Confining as these rules of thumb are, they still do not rule out all answer keys that key balancers judge unacceptable, such as the cyclic:

A B C D A B C D A B C D A B C D A B C D A B C D A

or the palindromic:

A A B B B C C C D D D A D A D D D C C C B B B A A

Other informal guidelines that might be added to the list above are:

4. Do not exclude runs altogether, have some short ones (e.g., at least one run of three and two runs of two).

5. Avoid overly patterned sequences, such as obvious symmetries or repeated cycles.

A loose, but comprehensive, rule of thumb for local balancing is: "Just make the key *look* random." Using a Monte Carlo simulation of 100,000 random sequences, we found that in a section of 25 four-choice questions, the percentage of acceptable

Maya Bar-Hillel is Professor, The Center for Rationality, The Hebrew University, Jerusalem, Israel (E-mail: maya@huji.ac.il). Yigal Attali is Associate Research Scientist, The Educational Testing Service, Princeton, NJ 08541. The second author was at Israel's National Institute for Testing and Evaluation at the time this research was conducted. We thank NITE for its support. Without access to their data much of this work could not have been carried out.

DOI: 10.1198/000313002623

(i.e., properly key balanced according to Rules 1–4.) sequences is around 17%. Hence, "delicate art."

Jessell and Sullins (1975, p. 45) advocated an even more restrictive form of key-balancing, talking of "an ideal format with each option as the keyed response for one-fourth of the items and with the keyed response position appearing no more than *twice* [italics ours] in sequence." Taken literally, this recommendation leaves but 3% of all possible 25-long, 4-option sequences.

Considering how widespread key balancing is, it is surprising how little the psychometric literature has to say about it. In a recent survey of "46 authoritative textbooks and other sources in the educational measurement literature," Haladyna and Downing (1989a, p. 37) found 38 that addressed the issue of key balancing, with all but one recommending it. Yet this recommendation is supported by neither data nor theory. Indeed, Millman and Greene (1989) denied that anything but common sense is required to support key balancing: "Some rules, like [key balancing] ... make sense regardless of the outcome of empirical studies on [its] violation" (p. 353). The positioning of answer options has been all but ignored as a psychometric characteristic of interest, much less as one that could affect psychometric indices [for a survey of within-item positioning effects see Attali and Bar-Hillel (in press)]. As far as we can tell, even popular guides to passing multiple-choice tests ignore the topic [e.g., Barron's guide to the SAT by Brownstein and Weiner (1982); *The Princeton Review*'s guide to the GMAT by Martz and Katzman (1991)].

Reading the meager literature on key balancing, it is remarkable that the idea of randomization is hardly mentioned (but see Anderson 1952 and Mosier and Price 1945). It seems that key balancing is erroneously seen by some as synonymous with randomization.

## 2. WHAT IS THE ETS'S ANSWER-KEY POLICY?

NITE's present key balancing policy—"leave the balancing up to chance"—clearly, and elegantly, requires no secrecy. Because departures from randomization leave detectable traces, ETS's corresponding policy, even though not publicly admitted, can be extracted from the answer keys of their published tests.

Ten Real SATs appear in a book by that name (Claman 1997) produced by The College Examination Board. Each SAT test includes 128 multiple-choice questions distributed over six sections that vary in length, typically: 10, 13, 15, 25, 30, and 35 items. Each question has 5 options (except the section with 15 questions, with 4 options for each question). Thus, the 10 tests in the book included a total of 1,280 questions, 1,130 of which were five-option questions. SAT keys seem to have been balanced similarly to PET keys. We found that: (1) In the 25-question sections, all positions appeared between 3 and 8 times, inclusive; in the 30-question sections all positions appeared between 4 and 9 times; in the 35-question sections all positions appeared between 5 and 11 times. (2) Correct answers were never placed more than three times in a row in the same position. (3) In the shorter sections (10, 13, and 15 items—comparable in length to half a PET section) correct answers occupied all positions. Can this similarity to NITE's policy be just coincidence? We show that these properties are unlikely to be generated at random.

*Table 1. Observed Versus Expected Frequencies of Runs in SAT Answer Keys*

| *Run length* | *1* | *2* | *3* | *4* | *5* | *6* | *>6* |
|---|---|---|---|---|---|---|---|
| Observed in 10 real SAT tests | 916 | 152 | 20 | 0 | 0 | 0 | 0 |
| Expected by chance | 828 | 162 | 32 | 6 | 1.2 | 0.2 | 0.06 |

1. The SAT's answer keys are not just balanced, they are overly balanced: in other words, they exhibit less variance around perfect balancing than would be expected by chance. To show this, we computed the absolute difference between the expected number of correct answers in the extreme positions and the observed number in an SAT section. For example, in a section of 25 questions, if 8 correct answers are in position A or E, rather than the expected 10, the absolute difference is 2. We then summed these absolute differences over the SAT's six sections. To obtain the expected distribution of these sums, we performed a Monte Carlo simulation of 100,000 randomly positioned SAT answer keys, and calculated them for each simulated key. The mean was 10.5 (SD = 3.4), and the median was 10.3. The observed sums for each of the 10 real SATs had percentile ranks in the simulated distribution of 4, 4, 7, 7, 13, 17, 25, 32, 47, and 75, respectively. Note that 9 of the 10 SAD's have a percentile rank less than 50 ($p$ = .01, sign test), and the median percentile rank is a low 15.

2. There are too few long runs in the SAT's answer key: To get the number of runs of varying length that would be expected if correct answers were placed at random, we used a Monte Carlo simulation of 100,000 SAT-like sequences of correct answers. Table 1 shows the expected distribution of run lengths if correct answers were randomly positioned, as compared with the distribution of run lengths actually observed in the 10 SATs. The difference between them is significant (chi-square $= 57_{6\,\mathrm{df}}, p < .0001$). Hence, it is a safe guess that ETS has a policy of deliberately avoiding runs longer than three. Moreover, even runs of two and three are underrepresented.

This has implications for the key's repetition rate. For five-choice questions, the proportion of times that a correct answer was in the same position as in the preceding question was 16% (calculated over all 1,080 pairs of adjacent items in the $10 \times 5 = 50$ five-choice sections; $p$ = .0004), rather than the expected 20%.

3. The SAT's answer keys have too little position-neglect: We computed the expected number of times that the key to an SAT short section would miss one position altogether, if correct answers were positioned at random. In the section with 10 questions (and 5 options), the probability of such an event is 0.54; in the section with 12 questions (and 5 options), the probability of such an event is 0.34; and in the section with 15 questions (and 4 options), the probability of such an event is 0.05. But in all 30 (3 × 10) of the short sections considered, every position always appeared in the answer key at least once. The probability of this happening by chance is less than .0001. Hence it, too, seems to be a matter of policy, not coincidence.

We infer that ETS's "delicate art of key balancing" resembles NITE's now-defunct policy, requiring local as well as global balancing.

*Table 2. Effect of the Underdog Strategy on Probability of Correct Guessing, P(CG), and on SAT Score*

| *Ability level* | *P(CG) in long sections* | *P(CG) in short sections* | *P(CG) in entire test* | *P(CG) in verbal subtest* | *Gained points* | *Points' estimated worth* | *Gain in SAT points* |
|---|---|---|---|---|---|---|---|
| .9 | .31 | .38 | .33 | .32 | 1.2 | 11 | 13 |
| .8 | .29 | .36 | .31 | .30 | 1.9 | 7 | 13 |
| .7 | .28 | .34 | .30 | .28 | 2.3 | 7 | 16 |
| .6 | .27 | .32 | .28 | .27 | 2.6 | 5 | 13 |
| .5 | .25 | .30 | .27 | .26 | 2.7 | 5 | 14 |
| .4 | .24 | .29 | .26 | .24 | 2.6 | 6 | 16 |
| .3 | .24 | .27 | .25 | .23 | 2.4 | 6 | 14 |
| .2 | .23 | .25 | .23 | .22 | 1.9 | 7 | 13 |
| .1 | .21 | .24 | .22 | .21 | 1.2 | 9 | 10 |

## 3. HOW TO GUESS (IN MULTIPLE-CHOICE TESTS) IF YOU MUST

The main objection to balancing answer keys is that balanced keys can be exploited, enhancing the testwise test taker's chances of guessing correctly. The following is a simple strategy for such exploitation. We call it "The Underdog Strategy":

a. Answer all the questions in the section you can.

b. Count the frequency of each position among your answers.

c. Select the position with the lowest frequency—the "underdog" position (in case of a tie, any one of them will do).

d. Give the underdog position as the answer to all as-yet-unanswered questions.

We carried out a Monte Carlo to compute the benefit of this strategy in the SAT. 10,000 test takers, each of whom took all 10 SATs, were simulated at each of nine knowledge levels, from 10% to 90%. For each knowledge level, a percent of the questions exactly corresponding to that level were "answered correctly," at randomly chosen positions throughout the entire ten tests. Then the remaining questions within each section were "guessed" according to the Underdog strategy. For the Verbal subtest only, the mean proportion of successful guesses was translated into the number of points this strategy adds over the number expected from random guessing (or, equivalently, from omitting) under the SAT's scoring rule. (SATs are formula-scored by adding a point for each correct answer, subtracting 1/4 of a point for each error, and giving no points for omissions.) Table 2 shows the simulated mean impact of this strategy on an examinee's score.

The strategy's benefit relates to one's knowledge in two opposing ways. On the one hand, the more questions one knows, the larger the probability that an Underdog guess will be correct. This is shown by the monotonic increase in the proportion of correct guesses from low to high knowledge (columns 2, 3 and 4). Note in particular the dramatic effect in short sections (column 3), reflecting the fact that the shorter the window within which key balancing is practiced, the greater the potential benefit the Underdog bestows per item. To give an extreme example, if the key is perfectly balanced, then an examinee who knows the answer to all but one of the questions can simply deduce the position of the unknown answer. On the other hand, the more questions one needs to guess the answer to (namely, the fewer one knows), the more the question-by-question benefit of the Underdog accumulates. The net effect of these two opposing trends yields a nonmonotonic per-question advantage to the Underdog across knowledge levels (column 6).

Based on the Score Conversion Tables in Claman (1997), the number of SAT points that each question is worth also ranges nonmonotonically across knowledge levels, from about 10 in the extremes of the knowledge distribution, to about 5 for median knowledge (column 7). Thus, the added SAT points are not monotonic with increased knowledge (column 8, which is the product of columns 6 and 7). All told, the Underdog strategy can add between 10 and 16 points to one's Verbal SAT score, as compared with random guessing. A gain of this magnitude is about 50% higher than the estimated effect of coaching on scores for the Verbal subtest (Powers and Rock 1999)!

The Underdog strategy exploits a single feature of key balancing, namely global balancing. Clearly other features could also be exploited. For example, a strategy that exploits run avoidance is never to guess a position that repeats an adjacent position. A more sophisticated version of this strategy is never to guess a position that appeared in a window of adjacent positions. We checked several such strategies, and though all give the test taker a slight edge over chance, none does as well as the Underdog (Attali 2001, unpublished doctoral dissertation, The Hebrew University).

## 4. THE CASE FOR A RANDOMIZED KEY—AND AGAINST A BALANCED ONE

In a chapter devoted exclusively to advice on how to write multiple-choice test items, Haladyna (1994), after advising: "Balance the Key," had this to say: "... many test makers and experts on testing recommend that the correct answer for any test be evenly balanced among the response options.... Any serious departure from this rule may cause higher performing students to see patterns that may clue them toward guessing right answers and performing higher than they should perform. Or, low performers may randomly select a choice position, such as C, and accidentally get a higher than deserved score by pure luck" (p. 80).

The last sentence suggests that by "serious departure" Haladyna meant "a long run," such as a run of Cs. Indeed, as we noted at the beginning of this article, unattended keys are very likely to exhibit a preponderance of correct answers in middle positions, such as C (Attali and Bar-Hillel in press). Thus, Haladyna's concerns are valid if the alternative to balancing the key is to employ an answer key as is, since unattended keys are usually biased keys. But if the alternative to an unattended key is a randomized key, then the concerns are quaint, if not downright wrong. We consider them in turn.

Regarding the possibility of "seeing patterns": people can, and often do, "see patterns"—but they see patterns even where there aren't any. In unbalanced keys, the patterns test takers see may (or may not) reflect genuine biases and tendencies of the test makers, such as the middle bias. In balanced keys, the patterns seen may (or may not) be those imposed by the balancing policy, such as the absence of long runs. But in randomized keys, patterns can only exist ex post, hence even when "seen" they can never serve as valid clues for improved performance.

Regarding the possibility that guessers might "accidentally get a higher than deserved score by pure luck"—note that if the key is randomized, and the test is a $k$-choice test, the probability of a correct position-based guess, whether it is C or not, is precisely $1/k$. In other words, the probability of "accidentally [getting] a higher than deserved score by pure luck" is no higher for a randomly produced "pattern" than for a "non-pattern," or than that assumed by random guessing. It is $1/k^n$, where n is the length of the pattern, regardless of whether the test taker is guessing randomly, or is somehow position-biased, or follows some strategy or some intuitive positional "pattern" or "clue."

Another concern, which we have not found in print, but is often voiced in conversation, is that test takers might be misled into abandoning a correct answer when it happens to continue a run that seems to them too long. This possibility was tested, and turns out to be invalid (Bar-Hillel and Attali 2001): test takers hardly, if at all, shy away from a correct response merely because it continues a positional run. Moreover, in real tests, where they do not encounter runs longer than three in the correct key, they nonetheless produce them in their own answer sequences.

So key balancing seems to be an odd solution to a nonproblem. Even if test takers were thrown by the occasional long run, educating them that in random keys there are no sequential dependencies is a more proper response than catering to their erroneous expectations by substituting random-appearing keys for random keys.

The practice of key balancing has not hitherto been openly discussed by professional testing agencies, nor have we found any explicit exhortation to exploit it in lists of testwise strategies (e.g., Carter 1986; Haladyna and Downing 1989b; Millman, Bishop, and Ebel 1965), in popular guides to generic test taking (e.g., Duncalf 1994), or in preparation courses for the PET or the SAT. We suspect that this reflects a curious and inconsistent, but popular, belief that balanced keys cannot be exploited. Recall, for example, the earlier quote from Haladyna (1994). Jessell and Sullins (1975, p. 45) also justified key balancing on the grounds that balanced keys "avoid providing test takers with systematic devices which would enable them to 'beat' the test" (p. 45). Ironically, the opposite is, of course, the case: Whereas no person and no strategy can "beat" a randomized key, balanced keys can be exploited.

The confusion may derive from the ambiguity of the term "random key." A key may have been randomly generated ex ante, yet be grossly unbalanced or highly patterned ex post (see, e.g., Bar-Hillel and Wagenaar 1991). It may be as bad, or worse, for keys to be systematically unbalanced (e.g., exhibit a middle-bias) as it is for them to be systematically balanced. Nonetheless, imbalanced keys that accidentally result ex post from a randomization procedure, especially when the randomization is public knowledge, provide no clues to guessing right. It seems that when the psychometric literature advocated key balancing, it was implicitly comparing it to doing nothing. A more astute comparison would have been to compare key balancing to key randomization. Balancing, while superior to doing nothing, is inferior to randomizing.

Clearly, once the exploitability of key balancing is pointed out (e.g., via the Underdog strategy), it becomes part of the arsenal of testwiseness, and can no longer be ignored. This provides an important, indeed compelling, reason for immediate abandonment of the practice. Moreover, a policy of randomizing the answer key can be openly publicized, unlike a policy of key balancing (and the balancing rules themselves), usually kept as a professional secret, and thus accessible only to those who detect it. A balanced key (because it enhances the probability of a successful guess) creates, other things equal, an easier test than a randomized key, but its advantages are spread out unequally across knowledge levels (see Table 2), making it unfair and detracting from its validity.

A central theorem from game theory imparts a certain robust beauty to randomized keys, which also accounts for their transparency. Randomizing the key, for test makers, and randomizing the answer position, for guessing test takers, are equilibrium strategies. In other words, even if one "player" finds out that the other player is randomizing, neither has an incentive to depart from randomization unilaterally (e.g., Luce and Raiffa 1957). "OK," one might say, "so test makers should randomize. But can test takers really randomize in a real testing situation?" Reassuringly, it does not matter whether they do or not. Whatever they elect to do when guessing is equally good against a randomized answer key. The important thing is not in teaching, or instructing, test takers to randomize when guessing. The important thing is that they cannot do better, or, for that matter, worse, than randomize, no matter what they do. It follows, therefore, that they might as well concentrate on one question at a time. When answering any particular question, they need never look at the answers they gave to other questions, as they cannot benefit thereby. Under key balancing, test takers could look beyond the particular question they were struggling with, and benefit, however slightly, therefrom. We cannot rule out the possibility that some small fraction of test takers have been exploiting key balancing. For example, a colleague of ours, nowadays a Professor of Mathematics, took the Quantitative SAT some decades ago, and achieved a small measure of fame by scoring a perfect 800. "I did not actually know all the answers," he confided. "I had to guess some, and I used a variant of your Underdog strategy. . . ." It is probably the fact that most test takers seem not to have done so hitherto, that has allowed key balancing to survive.

*[Received March 2002. Revised September 2002.]*

## REFERENCES

Anderson, S. B. (1952), "Sequence in Multiple-Choice Item Options," *Journal of Educational Psychology*, 43, 364–368.

Attali, Y., and Bar-Hillel, M. (in press), "Guess Where: The Position of Correct Answers in Multiple-Choice Test Items as a Psychometric Variable," *Journal of Educational Measurement*.

Bar-Hillel, M., and Attali, Y. (2001), "Seek Whence: Answer Sequences and Their Consequences in Key-Balanced Multiple-Choice Tests," Discussion Paper 252 in www.ma.huji.ac.il/~ranb.

Bar-Hillel, M., and Wagenaar, W.A. (1991), "The Perception of Randomness," *Advances in Applied Mathematics*, 12, 428–454.

Brownstein, S. C., and Weiner, M. (1982), *Barron's How to Prepare for College Entrance Examinations: SAT* (11th ed.), Woodbury, NY: Barron's Educational Series.

Carter, K. (1986), "Test Wiseness for Teachers and Students," *Educational Measurement: Issues and Practices*, 5, 20–23.

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

Duncalf, B.(1994), *How to Pass Any Exam*, London: Kyle Cathie limited.

Haladyna, T. M. (1994), *Developing and Validating Multiple-Choice Test Items*, Hillsdale, NJ: L. Erlbaum Associates.

Haladyna, T. M., and Downing, S. M. (1989a), "A Taxonomy of Multiple-Chc Item-Writing Rules," *Applied Measurement in Education*, 2, 37–50.

——— (1989b), "Validity of a Taxonomy of Multiple-Choice Item-Writing Rules," *Applied Measurement in Education*, 2, 51–78.

Jessel, J. C., and Sullins, W. L. (1975), "The Effect of Keyed Response Sequencing of Multiple-Choice Items on Performance and Reliability," *Journal of Educational Measurement*, 12, 45–48.

Luce, R. D., and Raiffa, H.(1957), *Games and Decisions*, New York: Wiley.

Martz, G., and Katzman, J. (1991), *Cracking the System: The GMAT*, New York: Villard Books.

Millman, J., Bishop, C. H., and Ebel, R. (1965), "An Analysis of Test-Wiseness," *Educational and Psychological Measurement*, 25, 707–726.

Millman, J., and Greene, J. (1989), "The Specification and Development of Tests of Achievement and Ability," in *Educational Measurement* (3rd ed.), ed. R.L. Linn, New York: Macmillan, pp. 335–366.

Mosier, C. I., and Price, H. G. (1945), "The Arrangement of Choice in Multiple-Choice Item Options," *Educational and Psychological Measurement*, 5, 379–382.

Powers, D. E., and Rock, D. A. (1999), "Effects of Coaching on SAT I: Reasoning Test Scores," *Journal of Educational Measurement*, 36, 93–118.

# Seek Whence: Answer Sequences and Their Consequences in Key-Balanced Multiple-Choice Tests

Maya Bar-Hillel

Center for Rationality and Interactive Decision Theory

The Hebrew University of Jerusalem

&

Yigal Attali

National Institute for Testing and Evaluation

The Hebrew University of Jerusalem

June 2001

## Abstract

The professional producers of such wide-spread high-stakes tests as the SAT have a policy of balancing, rather than randomizing, the answer keys of their tests. Randomization yields answer keys that are, on average, balanced, whereas a policy of deliberate balancing assures this desirable feature not just on average, but in every test. This policy is a well-kept trade secret, and apparently has been successfully kept as such, since there is no evidence of any awareness on the part of test takers and the coaches that serve them that this is an exploitable feature of answer keys. However, balancing leaves an identifiable signature on answer keys, thus not only jeopardizing the secret, but also creating the opportunity for its exploitation. The present paper presents the evidence for key balancing, the traces this practice leaves in answer keys, and the ways in which testwise test takers can exploit them. We estimate that such test takers can add between 10 and 16 points to their final SAT score, on average, depending on their knowledge level. The secret now being out of the closet, the time has come for test makers to do the right thing, namely to randomize, not balance, their answer keys.

Keywords: key-balancing, randomization

This paper, and its previous companion paper (Attali & Bar-Hillel, 2001), explore the role of answer position in multiple-choice tests. Attali and Bar-Hillel (2001) showed strong and systematic within-item position effects in the behavior of both test takers and test makers -- even the professionals who produce the SAT, and they explored their psychometric consequences. The present paper deals with sequential (across-items) position effects, which are introduced primarily by test makers' ill-advised policy of key balancing.

## I. "The delicate art of key balancing", or: When randomization is too important to be trusted to chance.

Surprisingly, people writing a multiple-choice question tend to place the correct answer in a central position as much as up to 3 to 4 times as often as at an extreme position, apparently with little if any awareness of this tendency (Attali & Bar-Hillel, 2001). Banks of multiple-choice questions therefore usually exhibit a preponderance of answers in middle positions. If the correct answers are not reassigned to different positions, the resulting answer key could be heavily unbalanced. The near-universal method of dealing with this bias is through the so-called "delicate art of key balancing". Key balancing is not an openly practiced policy [1], and its secrecy is maintained for good reasons -- the same reasons that should have mitigated, as we shall see, against the very practice.

Key-balancing was, until recently, the unwritten answer key policy at NITE, Israel's National Institute of Testing and Evaluation. As a result of the present work, the practice was abandoned in 1999, in favor of key randomization, so we are now free to divulge its details.

NITE produces and administers the Psychometric Entrance Test (PET), which measures various scholastic abilities and achievements, and is used for student admissions by all Israeli universities. In many ways it resembles the SAT, developed for similar purposes by the US' Education Testing Service (ETS), but it is a 4-choice test (the SAT is mostly 5-choice) consisting of two subtests of size 25 (Quantitative), two of size 27 (Verbal), and two of size 30 (English). Regarding the subtests of 25 questions, for example, NITE's policy was: i. No position should appear in the subtest key more often than 9 times, or less often than 4. ii. Correct answers should never be placed over three times in a row in the same position. iii. A sequence of about a half the length of the subtest (i.e., about a dozen consecutive items) should not lack one of the 4 positions. The first can be called "global balancing", albeit at a subtest level, while the rest are more local balancing practices. Note that global balancing, once achieved, is independent of item reordering, whereas run avoidance and position-neglect avoidance, the local properties, depend on the actual sequencing of the questions.

Hence, all of the following answer keys, though most are globally balanced, and seemingly "random", would be ruled out:

A B C B A B B D C B A B D B A C C C A B C D C A A (only 3 Ds, violates i.)

A B C D A B B D C B A B D B C C C C A B D C C A A (run of 4 Cs, violates ii.)

A B B C B A A A C B C C C D D C A A D D D A B B D (no D in first half, violates iii.)

Confining as these rules of thumb are, they still do not rule out all answer keys that would probably be judged unacceptable, such as the cyclic:

A B C D A B C D A B C D A B C D A B C D A B C D A

or the palindrome:

A A B B B C C C D D D A D A D D D C C C B B B A A

A looser, but more comprehensive, rule of thumb for local balancing is: "Just make the key look random". Other informal guidelines that might be added to the list above suggest how that recommendation might be applied: iv. Have some runs (keeping them under 4 long), don't exclude them altogether. v. Avoid overly patterned sequences, such as obvious symmetries or repeated cycles. In a subtest of 15 4-choice questions, the percentage of acceptable (i.e., key balanced) sequences is less than 25%[2]. Hence: "delicate art". Similar guidelines applied to the other subtest lengths.

For some, key-balancing is even more restrictive. For example, Jessell and Sullins (1975, p. 45) talk of "an ideal format with each option as the keyed response for one-fourth of the items and with the keyed response position appearing no more than twice [italics ours] in sequence". Whether you are a professional test-writer, or just someone who occasionally writes tests, we invite you to ponder how well any of these considerations capture your own answer-key policy, insofar as you have one.

Considering how widespread key balancing is, it is surprising how little the psychometric literature has to say about it. In a recent survey of "46 authoritative textbooks and other sources in the educational measurement literature", Haladyna and Downing (1989a, p. 37) found that 38 of them addressed the issue of key-balancing, and all but one recommending it. Yet this recommendation is supported by neither data or theory. Indeed, Millman and Green (1989) deny that anything but common sense is required to support key balancing: "Some rules, like [key balancing] ... make sense regardless of the outcome of empirical studies on [its] violation" (p. 353). The positioning of answer options has been all but ignored as a psychometric characteristic of interest, much less as one that could affect the psychometric indices of either individual items or entire tests, and insofar as answer position has been addressed, only within-item positioning has received attention (see a survey in Attali & Bar-Hillel, 2001). As far as we can tell, even popular guides to passing multiple-choice tests ignore the topic (e.g., Barron's guide to the SAT by Brownstein & Weiner, 1982; The Princeton Review's guide to the GMAT by Martz & Katzman, 1991).

Reading the meager literature on key balancing, it is remarkable that the idea of randomization is hardly mentioned (but see Anderson, 1952 and Mosier & Price, 1945, who offer strategies for randomizing answer keys ). It seems that key-balancing is seen by some -- erroneously, to be sure -- as synonymous with randomization. Thus, Jessell and Sullins (1975, p. 45) say : "Nearly every basic educational measurement textbook ... usually [recommends] that the correct answer appear in each position about an equal number of times and that the items be arranged randomly", and a few lines later they indicate that in "an ideal format" the keyed response should not appear "more than twice in sequence". Clearly, in spite of citing Anderson (1952) and Mosier & Price (1945), Jessell and Sullins cannot be talking about real randomization, even if they think they are.

## II. What is ETS's answer-key policy?

NITE's present key-balancing policy -- leave the balancing up to chance -- clearly, and elegantly, requires no secrecy. However, we did not presume to ask ETS about their corresponding policy. Fortunately for us, however, and unfortunately for those who want to keep key balancing a trade secret, position policies leave traces through the properties of the answer keys in which they result. So in lieu of a direct question, we extracted ETS's answer key practices from the answer keys of their published tests.

Ten real SAT tests appear in a book by that name (Claman, 1997) put out by The College Examination Board. Each SAT test includes 128 multiple-choice questions distributed over 6 subtests that vary in length: 10, 13, 15, 25, 30 and 35 items. Each question has 5 options (except the subtest with 15 questions, which has 4 options for each question). Thus the 10 tests in the book included a total of 1280 questions, 1130 of which were 5-option questions. Even though the SAT is primarily a 5-choice test, whereas the PET is 4-choice, their keys seem to have been balanced similarly. We found that: i. In the 25-long subtests, all positions appeared between 3 and 8 times, inclusive, in the 30-long subtests all positions appeared between 4 and 9 times, in the 35-long subtests all positions appeared between 5 and 11 times. ii. Correct answers were never placed over three times in a row in the same position. iii. In the shorter subtests (10, 13 and 15 items -- comparable in length to half a PET subtest) correct answers occupied all positions. Can this similarity to NITE's policy be just coincidence? We put this to a statistical test, namely, we tested that such properties are unlikely to be randomly produced.

1. The evidence that the answer keys are overly balanced appears in Attali and Bar-Hillel (2001, Section V).

2. To get the number of runs of varying length that would be expected in 10 real SATs if correct answers were placed at random, we used a Monte Carlo simulation of 60,000 SAT-like sequences of correct answers (a combinatorial calculation was complicated by the variable lengths of the SAT subtests, and the fact that some are not 5-choice). Table 1 shows the expected distribution of run lengths in random positioning of correct answers, as compared with the distribution of run lengths actually observed in the 10 real SAT tests. The difference between the expected distribution and the observed one is significant (chi-square=$57_{6df}$, p<.0001). It is a safe guess that ETS shares NITE's policy of deliberately avoiding runs longer than 3.

In addition to total avoidance of runs longer than three, shorter runs -- 2 and 3 -- are also underrepresented in these SAT tests. Consider the proportion of times that a correct answer is in the same position as in the preceding question. For 5-choice questions the expected proportion is obviously 20%, but in the 10 SAT Tests it was 16% (calculated over all 1080 pairs of adjacent items in the 10x5=50 5-choice subtests; p=.0004).

**Table 1**

**Observed and expected frequency of answers in various run lengths in SAT answer keys**

| Run length | 1 | 2 | 3 | 4 | 5 | 6 | >6 |
|---|---|---|---|---|---|---|---|
| Observed in 10 real SAT tests | 916 | 304 | 60 | 0 | 0 | 0 | 0 |
| Expected by chance | 827 | 324 | 95 | 25 | 6 | 1.5 | .42 |

3. What is the expected number of times that the key to an SAT short subtest would miss one position altogether, if correct answers were positioned at random? In the subtest with 10 questions (and 5 options), the probability of such an event is 54%; in the subtest with 12 questions (and 5 options), the probability of such an event is 34%; and in the subtest with 15 questions (and 4 options), the probability of such an event is 5%. But in all 30 (3 x 10) of these short subtests, every position always appeared in the answer key at least once. The probability of this is less than .0001. Hence this, too, seems to be a matter of deliberate policy, not coincidence.

We infer that ETS's delicate art of key balancing resembles NITE's defunct policy, namely: i. placing the correct option in roughly equal proportions in the possible positions over the entire subtest; ii. avoiding runs longer than three; and iii. giving all positions representation, even in "windows" as short as about a dozen items.

## III. How to guess (in multiple-choice tests) if you must.

The main reason for not balancing answer keys is that balanced keys can be exploited, enhancing the testwise test taker's chances of guessing correctly. The following is a simple strategy that a guessing examinee taking the SAT would benefit from using. We call it "The Underdog Strategy" :

a. Answer all the questions in the subtest you can.

b. Count the frequency of each position among your (hopefully correct) answers.

c. Select the position with the lowest frequency (the "underdog" position). If two or more positions are tied for underdogs, select any one of them.

d. Give the underdog position as the answer to all as yet unanswered questions [3].

We carried out a Monte Carlo simulation to compute the benefit of this strategy in the SAT using the same 10 real SAT tests mentioned in the previous section. The performance of 10,000 test takers, each of whom takes all 10 SAT tests, was simulated at each of nine knowledge levels, from 10% to 90%. For each knowledge level, a percent of the questions exactly corresponding to that level were "answered correctly", at randomly chosen positions throughout the entire ten tests. Then the remaining questions in each subtest were "guessed" according to that test's Underdog strategy. For the Verbal subtests only, the mean proportion of successful guesses was translated into the number of points this strategy adds over the number expected from random guessing (or, equivalently, from omitting) under the SAT's formula scoring. SATs are scored by adding a point for each correct answer, subtracting 1/4 of a point for each incorrect answer, and giving no points for omissions. We did not do a similar calculation for the Quantitative subtests, because the Quantitative score is partly based on open-ended questions, which complicates matters. Table 2 shows the simulated mean impact of this strategy on an examinee's score.

**Table 2**

**Effect of the Underdog Strategy on Probability of Correct Guessing, P(CG), and on SAT Score**

| Ability Level | P(CG) in Long Subtests | P(CG) in Short Subtests | P(CG) in Entire Test | P(CG) in Verbal Subtests | Gained Points | Points' Estimated Worth | Gain in SAT Points |
|---|---|---|---|---|---|---|---|
| .9 | .31 | .38 | .33 | .32 | 1.2 | 11 | 13 |
| .8 | .29 | .36 | .31 | .30 | 1.9 | 7 | 13 |
| .7 | .28 | .34 | .30 | .28 | 2.3 | 7 | 16 |
| .6 | .27 | .32 | .28 | .27 | 2.6 | 5 | 13 |
| .5 | .25 | .30 | .27 | .26 | 2.7 | 5 | 14 |
| .4 | .24 | .29 | .26 | .24 | 2.6 | 6 | 16 |
| .3 | .24 | .27 | .25 | .23 | 2.4 | 6 | 14 |
| .2 | .23 | .25 | .23 | .22 | 1.9 | 7 | 13 |
| .1 | .21 | .24 | .22 | .21 | 1.2 | 9 | 10 |

The strategy's benefit relates to one's knowledge in two opposing ways. On the one hand, the more questions one knows, the larger the probability that an Underdog guess will be correct. This is shown by the monotonic increase in the proportion of correct guesses from low to high knowledge (columns 2,3 and 4). Note in particular the dramatic effect in short subtests (column 3), reflecting the fact that the shorter the window within which key balancing is practiced, the greater the potential benefit the Underdog bestows per item. To give an extreme example, if the subtest consists of only 5 questions, and the key is perfectly balanced, then an examinee who knows the answer to four of the questions can simply deduce the position of the fifth answer. On the other hand, the less one knows, namely, the more questions one needs to guess, the more opportunity the question-by-question benefit of the Underdog over random guessing accumulates. The net effect of these two opposing trends yields a non-monotonic per-question benefit to the Underdog across knowledge levels (column 6).

Based on the ten Score Conversion Tables in Claman (1997), the number of SAT points that each question is worth also ranges non-monotonically across knowledge levels, from about 10, in the extremes of the knowledge distribution, to about 5, for median knowledge (column 7). Thus, the added SAT points are not monotonic with increased knowledge level (column 8, which is the product of columns 6 and 7).

All told, the Underdog strategy can add between 10 and 16 points to one's Verbal SAT score, as compared with random guessing. This might be an underestimate of its benefit for the entire SAT, because the Verbal subtests are on average longer than the Quantitative subtests, and the shorter the subtest, the greater the benefit (compare columns 2 and 3; 4 and 5). This gain is about 50% higher than the effect that coaching is estimated to have on scores for the SAT's verbal section (Powers & Rock, 1999).

The Underdog strategy's benefit is entirely due to exploiting a single feature of key balancing, namely global balancing (feature i). Clearly other features could also be exploited. Thanks to the negative dependency between items created by global balancing, the least frequent position among one's correct answers really is, on average, more likely than chance to be correct for the as-yet unanswered items. Step d. advocates that the Underdog position be consistently given to all the guessed items. Unintuitive as this might seem, it enjoys the same advantage that consistent prediction of the more probable event has over probability matching, in experiments by that name (e.g., Estes, 1976).

The Underdog strategy cannot be implemented in adaptive testing, where one cannot return to a previously unanswered question, and where the score is not a linear transformation of number of questions answered. But insofar as runs are avoided when selecting questions in adaptive testing too (clearly another trade secret, and one whose traces are more elusive than in printed tests), there might be a benefit to over alternation in guessing answers.

"But does not randomization also roughly result in global balancing?", one might ask. If the test is long enough, it surely might (though only probabilistically). Yet only imposed balancing has an exploitable built-in negative dependency. The kind of balancing that results from randomization in the long run (Law of Large Numbers) is inherently impervious to exploitation. To believe otherwise is to exhibit the notorious Gambler's Fallacy. Thus, replacing key balancing by randomization with some high probability will result in a balanced key -- but without the drawbacks (i.e., exploitability) of artificial balancing.

Almost as remarkable as the fact that professional test makers such as ETS have deliberately adopted exploitable policies of designing answer keys is the fact that these opportunities have apparently not been explicitly discovered or rampantly exploited. In that sense, they got away with their ill-advised policy. Perhaps that is why they have persisted till now. We shall return to this claim, and support it with evidence, in Section V.

## IV. Should test takers ever (attempt to) randomize?

The standard assumption in psychometric theory about the behavior of guessing examinees is that under complete uncertainty, they choose among the options at random (Lord & Novick, 1968). In contrast, we suggest that "guessing" should not be defined in terms of choice probabilities, because this would assume that which we wish to challenge. Rather, we suggest taking "guessing" to be that state of mind in which one must choose among options that one can find no good reason to choose among (by "options" we refer to the answers themselves, as distinguished from the positions which they occupy in an item's final form). It is widely acknowledged that Lord and Novick's assumption is overly simplistic (e.g., Budescu & Bar-Hillel, 1993; Attali & Bar-Hillel, 2001), but it has never been challenged in terms of where the options are positioned.

We propose that examinees regard every question as a problem solving task, in which their intent is to maximize their subjective probability of answering correctly (or, under formula scoring, their expected score). When they find themselves hard pressed to choose among the options on the basis of content, some examinees believe that they can nonetheless improve their probability of a correct guess by taking other considerations into account (this is commonly known as "testwiseness"). As the Underdog strategy shows, this belief is not unfounded -- exploiting sequential position effects in key balanced tests can indeed improve one's score[4]. Randomization, on the other hand, fixes the probability of answering correctly at 1/k, where k is the number of answer choices. Additionally, the Underdog strategy is simpler to apply than randomization. Once the known questions have been answered, mental effort is involved only in step b, counting. In contrast, randomization necessitates the actual operation of a random device over and over again for each guess anew. A random device must be used, because people have a faulty intuitive notion of randomness, which interferes with their ability to act as mental random devices (e.g., Bar-Hillel & Wagenaar, 1991; Falk & Konold, 1997; Rapoport & Budescu, 1997).

Thus, the randomization hypothesis is untenable on motivational, strategic, and psychological grounds. Its domination in psychometric models of guessing examinees is doubly ironic in light of the fact that the test makers themselves do not randomize the positioning of correct options.

## V. Do test takers (attempt to) balance their answer sequences?

We have not found any exhortation to exploit key balancing in popular guides to generic test taking (e.g., Duncalf, 1994), or in lists of testwise strategies (e.g., Carter, 1986; Haladyna & Downing, 1989b; Millman, Bishop & Ebel, 1965), or in preparation courses for the PET or the SAT. We suspect that this is a result of the curious, and inconsistent, belief that balanced keys cannot be exploited. For example, Jessell and Sullins (1975, p. 45) justify key balancing on the grounds that balanced keys "avoid providing test takers with systematic devices which would enable them to "beat" the test" (1975, p. 45). Haladyna (1994, p. 80) justifies a balanced key by stating that "any serious departure from [it] may cause higher performing students to see patterns that may clue them toward guessing right answers and performing higher than they should perform". The exact opposite is, of course, the case: only a balanced key can be exploited by higher performing students. No person and no strategy can "beat" a randomized key!

Nonetheless, it might still be the case that test takers spontaneously, and perhaps with little awareness, strive to produce a balanced response key. We next show that insofar as they do, the effect is too small to show up in the evidence of aggregate test taking behavior. In other words, for all practical purposes, key balancing on the part of test takers is hardly an issue.

Since test makers exercise total control over the answer key, it is safe to conclude that their keys look just as they wish them to look. In contrast, test takers are not free to endow their answer sequences with the properties of their choice, insofar as their primary goal is to reproduce the test maker's answer key as closely as possible. Obviously, high ability examinees produce answer sequences that closely mimic the answer key -- that, after all, is what being "high ability" means -- and that are therefore about as locally balanced as we know the answer key to be. So finding a "delicately balanced" answer sheet for a high ability test taker is not diagnostic of a tendency to deliberately bring about such an answer key. But low ability examinees produce sequences that differ considerably from the answer key -- that, after all, is what being "low ability" means. Do their answer sequences nonetheless exhibit local balancing?

In our companion paper (Attali & Bar-Hillel, 2001), we show that the answer sequences of test takers taking the PET have a small preponderance of responses in middle positions (about 53%), at the expense of the extreme ones (about 47%). In other words, the answer keys they produce are not even globally balanced. Moreover, the higher the proportion of their guessed answers, the greater the preponderance of middle options in their answers (central positions are chosen in about 57% of guesses). This means that the greater the opportunity for test takers to exercise balancing (i.e., the more they are guessing), the greater the imbalance in their answer sequences. But in the present paper, we shall concentrate on local balancing only, namely on sequential dependencies between adjacent questions, such as too-short runs and too-high alternation rates.

The major difficulty in uncovering the positional strategies of test takers is that the scope of these strategies is naturally limited to questions which are guessed, and it is not always straightforward to identify those. Nevertheless, several approaches can be taken to this issue. One is to concentrate on test takers that can be safely assumed to be largely guessing. Another is to concentrate on erroneous answers, which can be safely assumed to be usually the result of guessing (we use "guessing" in opposition to "knowing", ignoring its many subtle varieties). We used both approaches.

### *Evidence from low ability test takers*

We arbitrarily took a sample of five recent PET test results from the data banks of NITE, concentrating only on the two Quantitative subtests (25 questions each). To enable us to assume that we were looking at answers that were largely the result of guessing, we concentrated on examinees who answered correctly 14 or less out of the total of 50 4-choice questions. These 551 examinees were in the two lowest percentiles. On average, this group was performing at no better than the chance expectation of 25% correct answers (12-13). Even the highest scorers among them were barely exceeding chance performance.

**Table 3**

**Observed and expected mean frequency of answer-runs of different lengths, in tests of 25 questions**

| Run length | 1 | 2 | 3 | 4 | 5 | 6 | >6 |
|---|---|---|---|---|---|---|---|
| Observed in the PET key | 17 | 3.3 | .50 | 0 | 0 | 0 | 0 |
| Expected by randomization (p=.25) | 14 | 3.5 | .83 | .20 | .05 | .01 | .004 |
| Observed in examinees' responses | 16 | 3.3 | .67 | .15 | .04 | .01 | .006 |
| Expected by randomization (p=.22) | 16 | 3.3 | .69 | .14 | .03 | .01 | .001 |
| Expected by randomization (p=.18) | 17 | 3.0 | .50 | .09 | .02 | .002 | .001 |

Table 3 shows the distribution of run lengths in the PET's answer keys (based on 10 subtests of 25 questions), and in these test takers' 1102 (551x 2) answer sequences (normalized to 10 x 25 = 250). First, Table 3 replicates with respect to the PET key what Table 1 showed regarding the SAT key: the PET's key is significantly unlike a randomized key (compare the numbers in rows 1 and 2, multiplied by 10 tests to get the frequencies; runs of length 4 and up were combined; chi-square=$8.3_{3df}$, p=.04). It has too many alternations (i.e., 169 runs of 1, as against 144 expected), too few runs of 2 and 3 (38 as against 43), and no runs of 4 or more (as against an expectation of 2.6).

The answer sequences that are produced by examinees who, due to low ability, must be largely guessing, differ not nearly as much from random sequencing (even though this difference, based as it is on an N of 1102, is statistically significant; compare rows 2 and 3, multiplied by 1102; chi-square=$150_{6df}$, p<.0001). Most notably, the examinees do not shy away from long runs (i.e., 4 or more) -- hardly more so than a random device does (.21 as against .26).

We calculated our examinees' alternation rate, namely, the proportion of times that their answer did not repeat the position of the immediately preceding answer. It was 78%. This is somewhat higher than the chance expectation of a 75% alternation rate (n=1102 x 24, p<.0001) -- but considerably lower than the 82% alternation rate of the correct-answer key. Note that the observed distribution (Table 3, row 3) is nicely approximated by a random process with a repeat probability matching that of the examinees, 22% (row 4). Although this repeat probability was calculated from our examinees' answers, the similarity between rows 4 and 3 is not a trivial result, because row 4 is based only on successive answers, whereas the correct key obviously has dependencies that go at least three deep (e.g., no runs longer than 3).

To summarize Table 3, low-ability guessing subjects show little evidence of run avoidance. Any departures they exhibit from random sequencing with a repeat rate of 25% can be more than accounted for by random sequencing with a repeat rate of 22%. The repeat rate in a balanced key is 18%. So any effect of run avoidance which might exist is quite small.

In spite of the nice match between the observed responses and what would be expected by randomization, it is unlikely, for the reasons spelled out earlier, that examinees of any ability, much less low ability, are deliberately randomizing. More likely, the apparent randomization is due to the cognitive strain under which they are answering, which depletes the resources required to deliberately create any sequential dependency (negative, in this case), such as by keeping a running tally of position frequencies.

One of the rare contexts where people have been previously observed to generate response sequences that satisfy standard tests of randomness reasonably well was identified by Rapoport and Budescu (1992) in a strictly competitive zero-sum game, where a strategy of randomization is optimal. Rapoport and Budescu concede that "We have obtained better results [than most studies on generating random sequences] ... because ... the ... task of monitoring the opponent's moves interferes with the subject's memory of past moves" (p.

362). This account is even more plausible here, since randomization is actually not the optimal strategy for test takers -- global balancing is better.

*Evidence from high ability test takers*

It is possible -- indeed, intuitively plausible -- that examinees really do wish to balance their answers, but that those of low ability are hardly in an ideal position to realize this desire, burdened as they are by the demands of the test. High ability examinees may be better positioned to worry about key balancing, but have less need, or opportunity, to exercise it, since their keys are roughly balanced as a side effect of their knowledge. Still, we aimed to study high ability examinees too.

Our data comes from the same 5 Quantitative tests that were analyzed earlier. We considered only examinees who, in 50 questions, erred at least once, so there would be something to look at, and no more than 5 times, so they are performing at a 90% knowledge level at least. We assumed that their errors are due to guessing. Table 4 presents the percent of times that a guess (or, more accurately, an erroneous answer) was in the same place as was the correct answer to the immediately preceding question, as a function of the examinee score (successive pairs of questions in which the correct answer was in the same position were obviously excluded from this analysis). In most cases, this percent is significantly less than the chance expectation of 33% (recall that we are only looking at questions where one of the distractors -- three in number -- were chosen, not a correct answer). However, the values are close to 30%, and average 31%, so the effect, though statistically significant, is very small -- as it was for the low-ability examinees.

**Table 4**

**Percent of times a wrong answer repeated the position of a previous correct answer**

| Raw Score | N of cases | Percent Repeat | p-value |
|---|---|---|---|
| 49 | 385 | 29 | .02 |
| 48 | 1178 | 30 | .01 |
| 47 | 1798 | 32 | .06 |
| 46 | 2756 | 31 | .01 |
| 45 | 3508 | 30 | .0001 |

*Questions-Without- Answers*

Although the test takers were not overly reluctant to repeat a previous position in a real test, they might still have a local aversion to repeat consecutive positions in a simplified situation of pure guessing. This was tested in the task which we named: Questions-Without-Answers. The respondents in this task were undergraduate students at The Hebrew University, who were approached in their classrooms at the end of a lesson, and asked to stay for a brief experiment during their break, in return for a prize of 200 NS that would be awarded by lottery to one of the participants, or as part of class requirements. A total of 141 respondents acquiesced. They were run in three groups -- students from Social Work, Education and Psychology -- of roughly equal size. They were requested to imagine that they were taking a multiple-choice test with four options. A first question was presented, with the position of the correct answer marked, but with no answers actually given, thus, for example:

Who was the first president of Israel?

A B C D*

Each of the positions was marked as correct for about one fourth of the respondents. It was followed by a second question, again given without answers, thus:

Who was the 17th president of the United States?

A B C D

Essentially, the respondents were requested to guess the position of the correct answer to a test question, knowing only the position of the correct answer to its predecessor. Only 12 of the 141 respondents repeated the position of their previous answer -- a third of the 25% expected under random choice (p<.0001).

Another group of 127 undergraduate Psychology students received the following two questions, in this order:

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

69 respondents "answered" both questions, and 58 others found the answer to the first question circled, and had to "answer" only the second question.  Details about the distribution of answers can be found in Attali and Bar-Hillel (2001, Section III).  For present purposes, we wish to note that only 10% of the respondents (6 in the pre-answered condition, and 7 in the self-answered condition) repeated the same answer position twice -- less than half of the 25% expected under random choice (p<.0001).

Putting together the results of real examinees taking the PET and those of the participants in the artificial Question-without-Answers task, we see that people may well wish to avoid runs when answering multiple-choice questions, and they alternate heavily in the artificial task, where nothing interferes with their ability to do so.  But in a real testing situation, they either are not concerned with runs, or lack the wherewithal to avoid them.  Perhaps the difficult items are too absorbing in themselves.  This can be true for low ability examinees as well as for high ability examinees.  For high ability examinees perhaps the concern is moot, because by and large they produce a balanced key insofar as by and large they are successful in reproducing the correct, hence balanced, key.

So even though test makers take pains to assure that the answer key of high stakes tests (such as the old PET) are free of runs longer than three, test takers show no corresponding bias against long runs.  They exhibit such runs in their answer sequences at rates approximating those which are expected by randomization, and certainly not at zero rates.

## VI.  Are test takers thrown by long runs in answer sequences?

Personal communication with NITE's test makers, and informal conversations with colleagues, raise another concern that seems to underlie key balancing.  Randomization might produce -- albeit, rarely -- sequences that would seem highly non-random to examinees, such as a long run of correct answers in the same position. A sequence like that could make the examinees doubt the correctness of their answers, to the point that they might alter them. The validity of this concern can be questioned both empirically and normatively. The normative question is whether test makers should let test takers' attitudes dictate their key policies. We address this in Section VII. The empirical question is what the test takers' attitudes really are.

Suppose that examinees are indeed disturbed by the appearance of a long run in their previous answers, such that when unsure of the correct answer to a present question they are reluctant to choose a position that would continue the run. If so, the percent of correct answers to a question which follows such a run would be lower than if that question were given in the absence of a previous run.

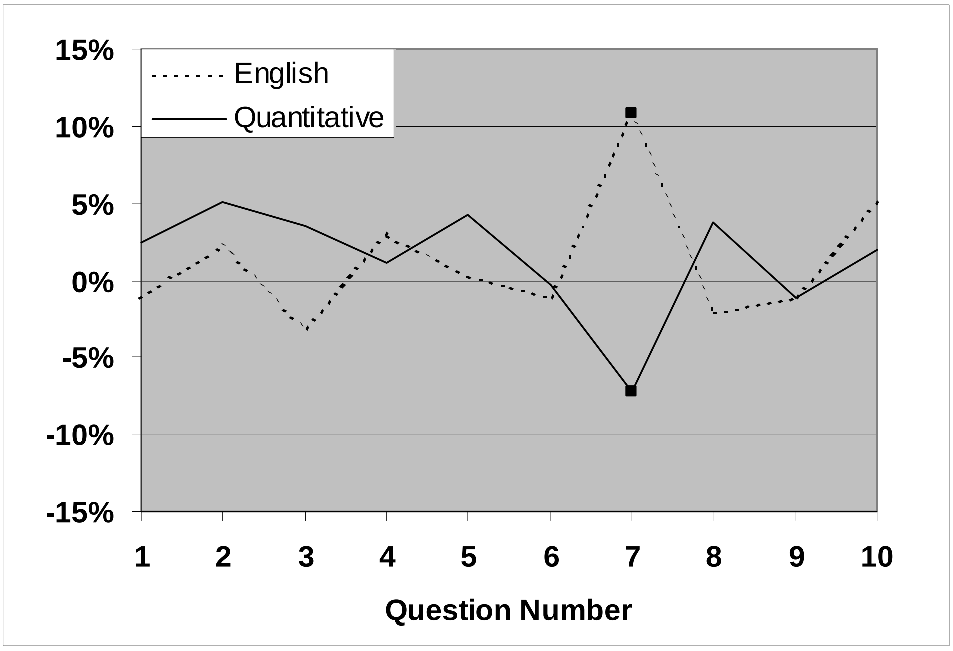
Jessell and Sullins (1975) gave test takers one of seven forms of a 60 item 4-choice test. Some received a test whose key had been balanced, in the large and in the small, and for the others, the key was altered to form a run of either seven or fourteen consecutive Bs. These were the first 7 (or 14) items, or the last 7 (or 14) items, or the middle 7 (or 14) items. Although the balanced test produced the highest proportion of correct responses -- 39%, the lowest performance, obtained with a run of 7 at the beginning, was, at 35.3%, not far behind, and the differences were not statistically significant. Jessell and Sullins concluded that their results "do not support the intuitively appealing notion that multiple-choice tests items should be keyed [in a balanced manner]" (p. 46), and that "examinees pay less attention to response patterning than might be supposed" (p. 47).

We conducted our own experiment. The PET exam has an experimental part which is not used for university selection. In that part, one PET Quantitative subtest and one PET English subtest were altered, by reordering the position of the correct answers so as to obtain an answer key in which the correct answers to the first seven questions were all in position B. No other change was made. The number of examinees taking these subtests was 202 and 995[5], respectively, for the original Quantitative and English subtests, and 278 and 235, respectively, for the altered Quantitative and English subtest.

Figure 1 plots the difference in percent correct between the original and the altered subtests, for each of the first 10 questions in these two subtests. If indeed guessing examinees are reluctant to choose the correct option B when it is preceded by a run of Bs, positive differences are expected in the first seven questions, because a smaller percent of the examinees would give a correct answer in the altered subtest, and negative differences are expected in the last three, because examinees who avoid B are now avoiding an incorrect option. But there is no indication of such tendencies in the results. The 2 black squares indicate the only differences, out of the 20 possible, whose 95% confidence intervals do not contain the null difference. In other words, the only difference that departed significantly from 0, obtained in the 7th item, was actually in the wrong direction – and it, too, is probably due to chance.

**Figure 1**

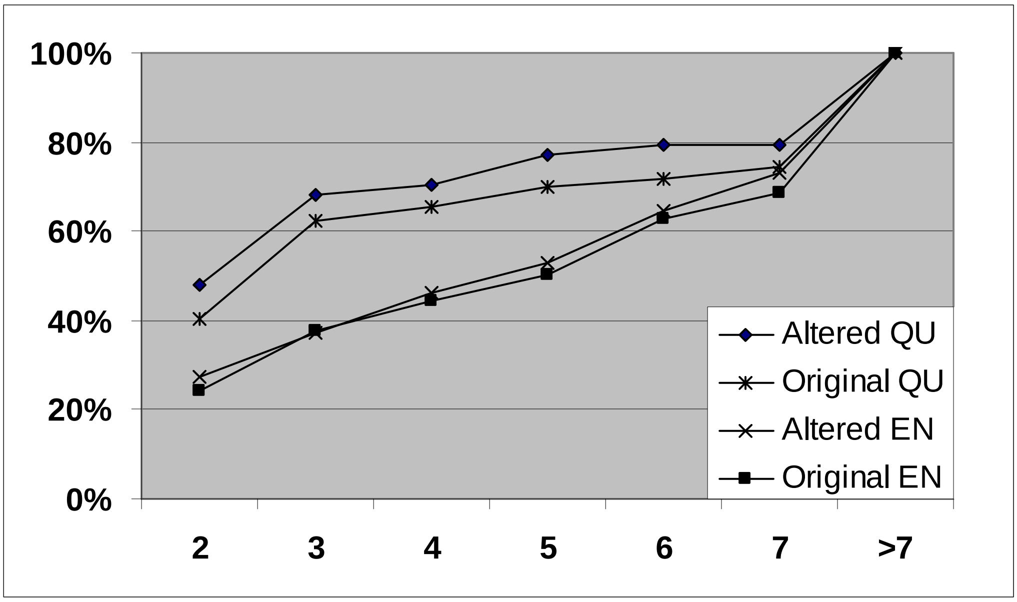
**Difference in Percent Correct Between Original and Altered Subtest**



Possibly the above analysis is not sufficiently sensitive, because not all the examinees actually encounter the long run in their answers, only those who answered them correctly. So Figure 2 plots the cumulative percent of errors only of those examinees who chose the correct answer in all preceding questions, namely those who erred for the first time on itme i, where i = 2, ...,7. The number of examinees who answered the first question correctly were: 180 (90%) for the original Quantitative subtest, 924 (93%) for the original English subtest, 241 (87%) for the altered Quantitative subtest, and 221 (94%) for the altered English subtest. Not surprisingly, the first question is equally difficult in the original as compared to the altered subtests, since it is preceded by a run in neither.

**Figure 2**

**Cumulative Distribution of First Wrong Answer**



Note that the two Quantitative subtests yielded a small difference in the expected direction, namely, the percent of (first) errors in the altered subtest is larger than in the original subtest. However this difference, 8%, is present as early as the second question and grows no larger up to the seventh question. The English subtests show even smaller, but also fairly constant, differences. These differences were not significant. Using the Kolmogorov-Smirnoff test for goodness-of-fit of two distributions, which is based on the maximal absolute difference, D, between the cumulative distributions, we could not reject the null hypothesis. A Chi-square with 2df (of $4D^2n_1n_2(n_{1+}n_2)$; see Siegel, 1956) yielded values of 2.4 (p=.31) for the Quantitative subtest and 1.5 (p=.47) for the English subtests.

## VII. The case for a randomized key -- and against a balanced one

In a chapter devoted exclusively to advice on how to write multiple-choice test items, Haladyna (1994), after advising: "Balance the Key", had this to say : "... many test makers and experts on testing recommend that the correct answer for any test be evenly balanced among the response options. ... Any serious departure from this rule may cause higher performing students to see patterns that may clue them toward guessing right answers and performing higher than they should perform. Or, low performers may randomly select a choice position, such as C, and accidentally get a higher than deserved score by pure luck" (p. 80).

These concerns are valid only if the alternative to balancing the key is to leave the key as it happens to be, in its natural, middle-heavy, state -- an inferior option. But if the alternative to a natural key is a randomized key, then the concerns are quaint, if not downright wrong.

Regarding the possibility of "seeing patterns", people can, and do, "see patterns", but they see them both where they are and where they are not (e.g., when they detect a non-existent "hot hand" in basketball, see Tversky & Gilovich, 1989). In unbalanced keys, the patterns they see may (or may not) reflect genuine biases and tendencies of the test makers, such as the one favoring middle positions. In balanced keys, the patterns seen may (or may not) be those imposed by the balancing policy, such as the absence of long runs. But in randomized keys, patterns can never be clues for improved performance.

Regarding the possibility that guessers might "accidentally get a higher than deserved score by pure luck" -- note that if the key is randomized, and the test is a k-choice test, the probability of a correct position-based guess (as contrasted with content or form based consideration), whether it is C or not, is precisely 1/k. Whatever strategy a test taker adopts for choosing an answer position to a guessed question, the probability that it contains the correct answer is precisely 1/k. In other words, the probability of "accidentally [getting] a higher than deserved score by pure luck" is no higher than that assumed by random guessing, even if the test taker is not guessing randomly at all, but rather is somehow position-biased, or follows some intuitive positional "pattern" or "clue".

Another concern, which we have not found in print, but was often voiced at NITE, for example, is that high-ability test takers might be misled into abandoning a correct answer when it happens to continue a run that seems to them too long. The results presented in Section VI belie this concern. Test takers hardly, if at all, shy away from a correct response merely because it continues a positional run. In real tests, where they do not encounter runs longer than 3, they nonetheless produce them. And in experimental tests, engineered so that they will encounter them (as in the experiment reported above where the key was reordered to produce a run of 7), they are hardly detracted by them. So key balancing seems to be an odd solution to a non-existent problem. But even had test takers been thrown by the occasional long run, the proper response is to educate them that in random keys there are no sequential dependencies, rather than to replace a random key with a random-appearing key, thereby catering to erroneous expectations.

Even though key balancing has not been openly acknowledged by professional testing agencies such as ETS and ACT (American College Testing) , and has not been explicitly recognized and exploited by preparatory courses and most test takers, once the exploitability of key balancing is pointed out (as in the Underdog strategy discussed above), it will no doubt become part of the arsenal of testwiseness. This is an important, indeed compelling, reason for their immediate abandonment. Moreover, once a policy of randomizing the answer key is adopted, it can be openly publicized, unlike a policy of key balancing (and the balancing rules themselves) which are usually kept as a professional secret, and thus accessible only to those who detect it. A balanced key (because it enhances the probability of a successful guess) creates, other things equal, an easier test than a randomized key, but its advantages are spread out unequally across knowledge levels, making it unfair.

A central theorem from game theory imparts a certain "robust beauty" to randomized keys, which also accounts for their transparency. Randomizing the key, for test makers, and randomizing the answer position, for guessing test takers, are equilibrium strategies. In other words, even if one "player" finds out that the other player is randomizing, neither has an incentive to depart from randomization unilaterally (e.g., Luce & Raiffa, 1957).

"OK, " one might say, "so test makers should randomize. But can test takers really randomize in a real testing situation?". The reassuring answer of the equilibrium theorem is that it doesn't matter whether they do or not. Whatever they elect to do when guessing is equally good against a randomized answer key. The important thing is not in teaching, or instructing, test takers to randomize when guessing. The important thing is in convincing them that faced with a randomized key, they cannot do better -- nor, for that matter, worse -- than randomize, no matter what they do. It follows, therefore, that they might as well concentrate on one question at a time. They need never look at their answers to other questions when answering any particular question, as they cannot benefit thereby. Under key balancing, test takers should have looked beyond the particular question they were struggling with, and could have benefited, however slightly, therefrom. Insofar as most test takers seem not to have done so hitherto, the testing establishment has gotten away with an embarrassing policy. It is high time that it now be abandoned.

**Conclusions**

Based on the literature on the perception of randomness, we expected to find a bias towards a balanced response sheet among test takers. In an ironical twist, we found that the bias is located squarely in the test makers' ball park, and it is introduced into answer keys through a deliberate, albeit misguided, policy, too-meticulously executed. In light of this, a corresponding tendency among test takers would not have been a bias, but rather a rational, adaptive response to the test makers' policy. Even so, little balancing is evident in test takers' response keys. Test takers either do not wish to have a balanced key, or just don't have the wherewithal to produce one in the stressful context of real tests. A randomized key, properly understood, would serve the test takers' best interest (holding test difficulty constant) [6] , by relieving them of the very need to balance. But most importantly, randomizing the key is doing the right thing.

Would "doing the right thing" by randomizing, rather than balancing, answer keys add to the validity and reliability of tests, such as the SAT? Considering that exploitation of the balanced key does not seem to have been rampant, the cautious answer would have to be "not much".  Still, such effects that this change of policy might have could only enhance, not detract, from these psychometric indices, because they would eliminate a source of variance that presently contributes to error variance, namely, whether the test taker does or does not attempt to exploit key balancing.   Proper instructions must very explicitly and forcefully drive it home that a test taker has absolutely nothing to gain by looking for, or noting, "patterns" in the answer sequence.  Such patterns, whether they exist or not, have absolutely no predictive value for the as yet unanswered questions.  The best that test takers can do is to devote all their attention and cognitive effort to the content of the questions and the offered answers -- nothing can be gained by spending any effort to consider answer positions -- neither within items nor across items.

In addition, once the possibility for exploiting a balanced key becomes public knowledge -- as, following this article, it no doubt will -- randomization becomes essential, to ward off the danger of a subsequent decline in validity.

Our conclusions and recommendations are equally valid for traditional paper-and pencil tests and for adaptive computerized tests.

## Notes

1. In April, 2000, while being a discussant of an earlier version of this paper, Linda Cook of ETS admitted that ETS practices "the delicate art of key balancing", whence the expression. The other discussant, Tim Davey of ACT, whose tests we did not analyze, only commented wryly that when he asked at ACT whether key balancing was being practiced, they refused to answer.

2. By Monte Carlo simulation, courtesy of Prof. Brendan McKay of the Australian National University.

3. For simplicity, we ignore situations of partial knowledge, for example one where the test taker may not know enough to select an answer to some question as in step a., but knows enough to eliminate an option options, including one that happens to occupy the underdog position. The Underdog strategy spelled out here cannot then be applied, but clearly a modified one can, such as answering with the "runner up" underdog, etc.

4. A colleague of ours, a Professor of Mathematics, told us that when he took the Quantitative SAT some decades ago at the age of 17, he achieved a measure of nationwide fame by scoring a perfect 800. "Only I knew", he confided, "that I did not actually know all the answers. I had to guess some, and I used a variant of your Underdog strategy".

5. There is no particular reason for the large number in one of the 4 groups -- it just so happened that this subtest was administered as part of larger number of whole tests.

6. If a test's questions and their answers are held constant, then a biased answer makes the test easier for the test taker, because the bias can be exploited to increase the probability of successful guessing. But if two tests with fixed answer keys are equally difficult, a balanced answer key makes the test taker's task easier, because she is spared the need to figure out how to exploit the non-randomness.

## References

Anderson, S.B. (1952). Sequence in multiple-choice item options. *Journal of Educational Psychology, 43*, 364-368.

Attali, Y., & Bar-Hillel, M. (2001). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Submitted for publication.*

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

Brownstein, S.C., & Weiner, M. (1982). *Barron's how to prepare for college entrance examinations: SAT (11th Ed.).* Woodbury, N.Y.: Barron's Educational Series.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 30,* 277-291.

Carter, K. (1986). Test wiseness for teachers and students. *Educational Measurement: Issues and Practices, 5*, 20-23.

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

Duncalf, B.(1994). *How to pass any exam*. London: Kyle Cathie limited.

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

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 14*, 31-318.

Haladyna, T.M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, N.J.: L. Erlbaum Associates.

Haladyna, T.M., & Downing, S.M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2,* 37-50.

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

Jessel, J.C., & Sullins, W.L. (1975). The effect of keyed response sequencing of multiple-choice items on performance and reliability. *Journal of Educational Measurement, 12,* 45-48.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luce, R.D. & Raiffa, H.(1957). *Games and decisions.* New York: Wiley.

Martz, G., & Katzman, J. (1991). *Cracking the system: The GMAT.* New York: Villard Books.

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

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: Macmillan.

Mosier, C.I., & Price, H.G. (1945). The arrangement of choice in multiple-choice item options. *Educational and Psychological Measurement, 5*, 379-382.

Powers, D.E. & Rock, D.A (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36*, 93-118.

Rapoport, A. & Budescu, D. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General, 121*, 352-363.

Rapoport, A. & Budescu, D. (1997). Randomization in individual choice behavior. *Psychological Review, 104*, 603-617.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Tversky, A. & Gilovich, T. (1989). The cold facts about the "hot hand" in basketball. *Chance, 2(1),* 31-37.