Mind & Society (2005) 4: 3–12
DOI 10.1007/s11299-005-0001-z

ORIGINAL ARTICLE

**Maya Bar-Hillel · David Budescu · Yigal Attali**

# Scoring and keying multiple choice tests: A case study in irrationality

Received: 5 July 2004 / Accepted: 11 October 2004


**Abstract** We offer a case-study in irrationality, showing that even in a high stakes context, intelligent and well trained professionals may adopt dominated practices. In multiple-choice tests one cannot distinguish lucky guesses from answers based on knowledge. Test-makers have dealt with this problem by lowering the incentive to guess, through penalizing errors (called *formula scoring*), and by eliminating various cues for outperforming random guessing (e.g., a preponderance of correct answers in middle positions), through *key balancing*. These policies, though widespread and intuitively appealing, are in fact "irrational", and are dominated by alternative solutions. *Number-right scoring* is superior to formula scoring, and *key randomization* is superior to key balancing. We suggest that these policies have persisted since all stake-holders – test-makers, test-takers and test-coaches – share the same faulty intuitions.

**Keywords** Formula scoring · Guessing · Key balancing · Multiple-choice tests · Randomization · Rationality · Testwiseness

## 1 Introduction

Over the last 35 years social and cognitive psychologists engaged in the experimental study of human judgment and decision processes have docu-

M. Bar-Hillel (✉)
Center for Rationality, The Hebrew University, Jerusalem 91904, Israel
E-mail: maya@math.huji.ac.it

D. Budescu
Department of Psychology, University of Illinois, Champaign, IL 61820, USA
E-mail: dbudescu@stat.psych.uiuc.edu

Y. Attali
Educational Testing Service, Rosedale Road, Princeton NJ 08541, USA
E-mail: yattali@ets.org

mented a wide variety of systematic violations of some basic tenets and implications of probability theory (e.g., Kahneman et al. 1982, Gilovich et al. 2002), and of classical decision theory (e.g., Kahneman and Tversky 2000). These robust empirical regularities are often cited as evidence of ''irrational'' behavior. Their conclusions were not universally embraced by all social scientists. Some economists downplayed their relevance, claiming that the small-scale demonstrations of irrationality would not survive in the ''real world'', where behavior is motivated by real and powerful incentives, and where people have ample learning opportunities (e.g., Grether and Plott 1979). Other critics argued that the problems used are often artificial, and are presented in unnatural and unrepresentative contexts. They would disappear, the claim went, if the problem were embedded in ecologically valid situations (e.g., Gigerenzer 1991).

The purpose of the present paper is to present a case study of irrationality in the familiar real-world context of large-scale, high-stake multiple-choice tests. We analyze and critique the practices with which the testing industry has, for decades, addressed the contentious issue of ''guessing'', and argue that these practices are as naïve and ''irrational'' as lay participant behavior in the psychologist's lab.

## 2 Scoring and keying multiple-choice tests

Multiple-choice tests enjoy many advantages that have made them tremendously popular tools in educational and psychological measurement. In the US, for example, by the time one graduates from high school, one has undergone numerous multiple choice tests administered by the state, the local schools, and as part of the college admissions process.[1] Multiple choice tests have been plagued since their inception with the so-called ''guessing problem'': a test-taker who does not know the correct answer to a question nonetheless has a non-negligible probability of selecting it by sheer luck.

Guessing (which here simply means ''answering with less than complete certainty'') can run the gamut from so-called ''wild'' or ''random'' guessing, where all options are chosen with equal probability, to partial knowledge or partial uncertainty, where the test-taker's probability of choosing some options might be higher or lower than that of choosing other options. In some cases, the probability of choosing some options can be as low as 0 (i.e., some options can be eliminated with certainty).

As a consequence of the guessing problem, the test-maker cannot distinguish between correct answers based on knowledge versus those deriving from a lucky guess. Hence, test-takers with different knowledge could end up with the same score, and test-takers with the same knowledge could end up with different scores. This state of affairs poses, *prima facie*, psychometric issues such as potential loss of validity, as well as ethical dilemmas related to the tests' ''fairness''. Test-makers have attempted to reduce the guessing

[1] According to US News and World Report (November 11, 2002), 2.2 Million US students take the SAT every year.

problem both by minimizing the incentives for guessing, and by reducing the opportunities for successful guessing.

*Minimizing the incentives* The simplest possible scoring rule for multiple-choice tests is to count the number of right answers, ignoring both omissions and errors (and hence also the distinction between the two). Under this scoring rule, *it always pays to answer*. More cautiously put, *it never pays to omit*. A guess is a gamble, but under number-right scoring it is a gain-only gamble, hence accepting the gamble dominates rejecting it. The test-makers' idea was to discourage guesses – at least those referred to as "wild guesses" – by penalizing errors, thus making the gamble embodied in guessing a risky one, where a test-taker could potentially lose points. The most popular scoring rule that this line of thought produced for scoring a k-choice item awards one point to a correct answer, but penalizes an erroneous answer to the tune of –1/ (k-1) points. Not answering – called an "omission" – scores 0 points. This solution was first suggested by Thurstone (1919).

*Minimizing the opportunities* A completely different approach to the guessing problem is to minimize the test-takers' probability of successful guesses. It is widely, and intuitively, assumed that when the test-taker has no idea what the correct answer is (or isn't), the probability of a successful guess is 1/k. But test-makers have long been aware of so-called "testwiseness" – the ability of test-takers to exploit some features of multiplechoice tests to enhance the probability of a successful guess above chance level. A common way of combating testwiseness is to identify cues that testwise examinees might exploit, and to take precautions against them. Thus, for example, test-makers are warned against an apparent tendency to make the correct answer longer, on average, than its distractors (perhaps as a result of the extra care invested in saying things just right), or against writing distractors that may be grammatically incompatible with the stem, etc. (e.g., Millman et al. 1965). Experienced professional test writers write questions that successfully remove many of these cues.

We will focus here on a positional cue, which, though widely believed to operate,[2] is seldom acknowledged explicitly: correct answers often appear disproportionately in middle positions. Indeed, a strong middle bias has been found in an assortment of tasks (e.g., Christenfeld 1995), including when writing or answering single multiple-choice questions (see Attali et al. 2003). Test-makers have addressed this bias by the practice known as *key balancing*, which involves making sure that, even in rather short sequences of questions, correct answers appear in each position a roughly equal number of times. Key balancing is an all but universal policy in professional settings, and may even be one of the hallmarks of professional test making. In a survey of "46 authoritative textbooks and other sources in the educational measurement literature", Haladyna and Downing (1989, p. 37) found 38 that addressed the issue of key balancing, with all but one recommending it.

[2] "... in Lake Wobegon, the correct answer is usually 'c'" (Keillor 1997, p. 180).

So far – so good. A problem was diagnosed – multiple-choice tests provide examinees with both the incentive and the opportunity to guess, thereby occasionally winning "undeserved" points. Two seemingly sensible and intuitively appealing "solutions" were proposed, respectively: lower the incentives for guessing by penalizing errors, and reduce opportunities for successful guessing by removing cues such as positional biases. In what follows, we take a critical look at these solutions and argue that they are naive and misguided. Although the solutions seem acceptable by lay standards, from the perspective of rational decision theory (game theory and expected utility theory) they look downright irrational. Indeed, we shall argue that the two practices are inferior even to their most obvious respective alternatives. Specifically, formula scoring is inferior to simple number-right scoring, actually compounding the guessing problem it was intended to reduce; and "the delicate art of key balancing", though superior to allowing the middle-bias to stand uncorrected, is inferior to key randomization, since it introduces another powerful cue to successful guessing.

## 2.1 Incentives: Does FS discourage guessing?

Formula scoring was algebraically designed to equate the expected score of a random guess with the sure-thing score of an omission – 0 points in both cases. Recall that the rationale behind it was to discourage guessing. Does it? This question can be answered from a normative-strategic point of view as well as from an empirical-psychological point of view, as follows.

In terms of expected score, answering dominates omitting, because it replaces a certain score of 0 by a subjectively expected score that is at least 0, but could be higher. If one believes oneself to possess some partial knowledge, then one's subjectively expected score from guessing is higher than 0. Also if one has no substantive knowledge, but is testwise, his or her subjectively expected score from "working the system" may be higher than 0. Only when resorting to "wild guessing" is one's expected score equal to 0. But wild guessing is rare; so, contrary to intent, formula scoring does not really discourage guessing, strategically speaking.

The above analysis notwithstanding, in point of fact formula scoring *does* discourage random guessing after all. Surprisingly, some people don't answer all questions even under the number-right scoring rule (e.g., data from the Educational Testing Service show a mean number of up to 3.5 unanswered items per test even in the GRE, which levels no penalty for erroneous guessing; M. Steffen, personal communication). Not surprisingly, formula scoring reduces guessing even further. First, many test-takers simply rely on the common, and misleading, formula scoring instructions that are deliberately worded to discourage guessing by failing to state that guessing is a (probabilistically) superior strategy (for an analysis of these instructions see Budescu and Bar-Hillel 1993). Second, formula scoring may discourage guessing in people who are risk averse, namely those who would reject a gamble in favor of its expected value. Such people might prefer the certainty of 0 points to a gamble between a higher or a lower score even where the gamble's expected value is higher than 0. (On the other hand, a minority of

risk seekers – those who prefer a gamble to its expected value – might systematically prefer guessing to omitting).

Be the reasons for test-takers' reluctance to answer under formula scoring what they may be, this reluctance is a psychometric Pyrrhic victory. Neither test-takers' risk attitudes nor their strategic reasoning are what multiple-choice tests purport to measure. From the perspective of educational measurement, individual differences in people's risk attitudes are nothing but unwanted noise: they add another source of variance, which is likely to reduce the test's reliability and validity. Furthermore, the test-makers' persisting inability to distinguish between a guessed correct answer and a known correct answer is compounded by their inability to distinguish between an omission deriving from ignorance and one deriving from risk aversion.

Table 1 compares formula scoring and number-right on several dimensions that are important and central to multiple-choice tests. It summarizes our arguments, showing that number-right is better than formula scoring not only on balance, but rather across the board.

Interestingly, formula scoring is used to score the Scholastic Aptitude Test (SAT) and the Graduate Record Examination (GRE) Subject tests, but number-right scoring is applied to the GRE General test, though these are all College Board tests. This inconsistency defies any attempt at rationalization. Apparently the original reasons for the particular scoring rules used by the different tests are rooted in some historical decisions or conventions. We speculate that they persist because transition from one scoring rule to another is deemed organizationally costly: just consider the need to re-calibrate the raw scores, and the discomfort of accounting publicly for the change.

The rule we presented for formula scoring is by far the most common, but we hasten to acknowledge that it is, of course, possible to design a scoring rule that would and should discourage guessing, even among the most avid risk-seekers. Indeed, for every confidence level P there exists some penalty function such that, unless one is more confident than P that one knows the correct answer, the expected score from guessing is lower than the score for omission. But the only way to discourage *all* guessing is by penalizing erroneous answers infinitely. And unless guessing is eliminated altogether, rather than just reduced in scope, the guessing problem does not entirely go away.

### 2.2 Opportunities: Does key balancing reduce the probability of a successful guess?

Key balancing is practiced to eliminate systematic "imbalances" and "patterns" in the answer key, which could be exploited by testwise test-takers to increase their probability of a successful guess. The essence of key balancing is to produce answer keys that "look random", namely, are locally representative (Kahneman and Tversky 1972; Rapoport and Budescu 1997). What "looks random" to test-makers mimics what lay participants in experimental studies produce when they are asked to randomize (Bar-Hillel and Wagenaar 1991), and is subject to the same biases. In the SAT, for example: (1) All

**Table 1** A comparison of number-right and formula scoring

| | NumberRight (NR) Scoring | Formula Scoring (FS) | Critique |
|---|---|---|---|
| Scoring procedure | Count number right only. | Count number right, number wrong, and omissions. | FS is more demanding. |
| Typical instructions | None, or: "Answer all questions". | "If you can eliminate some options, answer." | FS instructions are incomplete and misleading. |
| Ideal instructions | "Never-ever omit, you can only lose thereby." | "You may omit, but on average it doesn't pay." | FS's are self defeating. |
| Ethical issues | None; things are just as they seem. | Instructions which discourage guessing may undermine test-takers' best interest. | FS instructions are therefore unethical. |
| Underlying behavioral theory | Theory free. | Test-takers either know, can eliminate, or guess at random. | Theory underlying FS is overly simplistic. |
| Reliability and validity | Only source of error comes from knowledge. | Additional source of error from risk attitudes, and decision making. | FS introduces irrelevant variance. |

answer positions appear in the key of every subtest – even short ones – a roughly equal number of times; (2) There are no runs of same answer position longer than 3 in the answer key; (3) There are no windows in the answer key longer than 15 items with a missing answer position (Bar-Hillel and Attali 2002). Thus, through local representativeness, key balancing actually introduces into answer keys a powerful cue to successful guessing, namely, the negative dependencies between the positions of correct answers in successive questions.

Bar-Hillel and Attali (2002) simulated an easy to implement response strategy that exploits the negative dependencies induced by key balancing. According to this "underdog" rule, the test taker should first answer all the questions that he/she can. The guessed questions should then be given the answer occupying the position that appears least frequently in the answer sequence hitherto produced (the "underdog"). Compared with random guessing, this strategy adds between 10 and 16 SAT points on average to one's score (depending on one's knowledge level) – exceeding the best estimates of the marginal benefit of taking those costly, time-intensive coaching courses (Powers and Rock 1999)!

Table 2 compares key balancing with key randomization, on several relevant dimensions. The table indicates that there is not a single dimension on which key balancing is superior to key randomization.

## 3 Some implications for the persistence of formula scoring and key balancing

We do not pretend to know the psychometric price paid when using formula scoring rather than number-right, or key balancing rather than key randomization. Some studies have compared these alternatives, but many of them were plagued by major methodological problems (e.g., using inappropriate or inaccurate instructions for formula scoring), and their results are inconclusive. We are quite willing, however, to concede, however, that this price may be not too high.[3] Our focus is on other shortcomings of these practices.

Apart from its cumbersome implementation, the most serious cost of using key balancing is that it introduces a powerful test-wise clue. Hence, the policy (and the balancing rules themselves) must be kept as a professional secret. However, like any other test-wise clue, some test-takers eventually detect it and are able to exploit it. Moreover, the advantages from exploiting key balancing are distributed unevenly across knowledge levels (Bar-Hillel and Attali 2002), augmenting the unfairness and validity problems. Of course, the obvious and simple alternative policy of randomizing the answer key can be openly disclosed, because it is not exploitable in any way.

Fairness issues are associated also with the use of formula scoring. In addition to showing that formula scoring does not really solve the guessing problem, Budescu and Bar-Hillel (1993) showed that it introduces new

[3] Indeed, if true, this would actually explain the persistence of these two practices.

**Table 2** A comparison of key randomization and key balancing

| | Randomizing the answer key (KR) | Balancing the answer key (KB) | Conclusion |
|---|---|---|---|
| Procedure | Once formalized, can be mechanized. | Rule implementation, plus human judgment. | KR simpler, "cheaper". |
| Ideal instructions | "Every question stands alone. You have nothing to gain from looking elsewhere." | "You can do better than chance by balancing your answer key when possible." | KB can be exploited. |
| Transparency | Can, and should, be public knowledge. | Trade secret. | Only KR can be transparent. |
| Ethical issues | None; things are just as they seem. | Honest instructions cannot be given. | Hence KR + proper instructions is more ethical. |
| Underlying behavioral theory | None is required; whether test-takers appreciate randomness or not has no effect on score. | Random keys that don't appear random might throw off test-takers. | KB theory true only for some test-takers, not all. |
| Reliability and validity | Only source of error comes from knowledge. | Additional source of error from whether one tends to balance one's own key, too. | KB adds irrelevant score variance. |

problems not attendant in number-right scoring. The most important one is that it is very difficult for test-takers to figure out just the right decision strategy that this rule dictates. First, not all people can be trusted to calculate the expected score from wild guessing correctly, and to realize that it is the same as omitting. Second, not all people can be trusted to draw the proper normative implications of this fact, especially since these implications can vary from one situation to another (see some examples in Budescu and Bar-Hillel 1993), making it both unfair and unwise to leave it up to the examinees to draw the normatively appropriate conclusions from the scoring rule on their own.

But even when one's goal is simply to maximize expected score, it is not that easy to give the proper recommendations to those who cannot figure them out on their own. Nowhere is this difficulty more evident than in the dismal quality of the instructions given by those whose professional responsibility it is to get the instructions right. Most attempts to write proper instructions for examinees to guide them in making the right decision are flawed. For example, at one time or another, the College Board's site (www.collegeboard.com) has included examples of all the following inaccurate, incomplete and /or misleading recommendations: (1) Telling examinees that they will be penalized "a fraction of a point" for errors, *without stating the exact fraction*; (2) Telling examinees to "Omit questions that you really have no idea how to answer", *without indicating that the expected scores of omission and random guessing are equal* . (3) Telling examinees that "if ... you are able to eliminate one or more of the answer choices ... it may be to your advantage to answer", thus *implicitly inviting the inference that it does not pay to answer otherwise* . Since our analysis showed that, for an expected-score maximizer, even under formula scoring it is on average never better to omit than to guess, these are the very words that should have been in the instructions. But then, of course, these instructions would be self-defeating, as they would explicitly encourage always answering, when the *raison d'être* of formula scoring is to discourage guessing!

## 4 Final remarks

It is remarkable that two widely used "solutions" to the guessing problem that are inferior to two other readily available alternative practices have survived, almost unchallenged, for so long. The rules of key balancing mimic those employed by naive people attempting to "randomize": over-alternation, and short-term balancing that produce "locally representative" sequences. The rationale behind formula scoring also evokes lay logic in its simple-mindedness and myopia: "If you want to decrease guessing – penalize it".

Perhaps part of the reason for the persistence of these practices is the "tacit collusion" of all stakeholders in this process, based on their common intuitions and expectations about human behavior. For the most part, test-takers often omit items in formula-scored tests when they are in doubt about the correct answer. This sub-optimal strategy – induced, in part, by inaccurate instructions – is welcomed by many test-makers, who see it as

evidence that random guessing is suppressed. Coaching books, in turn, adopt the instructions given by the test-makers (''answer when you can eliminate an option'') uncritically, because they are intuitive. As for key balancing, an analysis of proper test-takers' behavior in response to it is much more difficult to perform, since the practice itself is a trade secret. Interestingly, the coaching industry did not devise simple ways to exploit this obvious weakness of key balancing. To the best of our knowledge, they were not aware of this practice, and they certainly failed to capitalize on its behavioral implications.

Our goal in this paper was to demonstrate that ''irrationality'' can exist and even flourish for decades in a high stakes, carefully designed and professionally controlled context, subject to scientific study and public scrutiny. Happily, alternative and less costly practices exist, ready to be implemented once the normative analysis is accepted as compelling.

## References

Attali Y, BarHillel M (2003) Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. Educ Measurement 40: 109–128

Bar-Hillel M, Attali Y (2002) Seek whence: Answer sequences and their consequences in keybalanced multiple-choice tests. Am Statistician 56: 299–303

Bar-Hillel M, Wagenaar WA (1991) The perception of randomness. Adv Appl Math 12: 428–454

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

Christenfeld N (1995) Choices from identical options. Psychol Sci 6: 50–55

Gigerenzer G (1991) How to make cognitive illusions disappear: Beyond heuristics and biases. In: Stroebe W, Hewstone M (eds) European review of social psychology, vol 2. Wiley, Chichester, UK, pp 83–115

Gilovich T, Griffin D, Kahneman D (eds) (2002) Heuristics and biases. Cambridge University Press, Cambridge

Grether D, Plott CR (1979) Economic theory of choice and the preference reversal phenomenon. Am Econ Rev 69: 623–638

Haladyna TM, Downing SM (1989) A taxonomy of multiple-choice item-writing rules. Appl Measurement Educ 2: 37–50

Kahneman D, Tversky A (1972) Representativeness: A study of subjective probability. Cognitive Psychol 3: 430–454

Kahneman D, Tversky A (2000) Choices, values and frames. Cambridge University Press, Cambridge

Kahneman D, Tversky A, Slovic P (1982) Judgment under uncertainty: Heuristics and biases. Cambridge University Press, Cambridge

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

Millman J, Bishop CH, Ebel R (1965) An analysis of testwiseness. Educ Psychol Measurement 25: 707–726

Powers DE, Rock DA (1999) Effects of coaching on SAT I: Reasoning test scores. Educ Measurement 36: 93–118

Rapoport A, Budescu D (1997) Randomization in individual choice behavior. Psychol Rev 104: 603–617

Thurstone LL (1919) A method for scoring tests. Psychol Bull 16: 235–240