To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring
Author(s): David Budescu and Maya Bar-Hillel
Source: *Journal of Educational Measurement*, Vol. 30, No. 4 (Winter, 1993), pp. 277-291
Published by: National Council on Measurement in Education
Stable URL: http://www.jstor.org/stable/1435226
Accessed: 18/07/2011 08:01

Journal of Educational Measurement
Winter 1993, Vol. 30, No. 4, pp. 277–291

To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring

David Budescu
University of Illinois, Champaign–Urbana
Maya Bar-Hillel
The Hebrew University, Jerusalem

Multiple-choice tests are often scored by formulas under which the respondent's expected score for an item is the same whether he or she omits it or guesses at random. Typically, these formulas are accompanied by instructions that discourage guessing. In this article, we look at test taking from the normative and descriptive perspectives of judgment and decision theory. We show that for a rational test taker, whose goal is the maximization of expected score, answering is either superior or equivalent to omitting—a fact which follows from the scoring formula. For test takers who are not fully rational, or have goals other than the maximization of expected score, it is very hard to give adequate formula scoring instructions, and even the recommendation to answer under partial knowledge is problematic (though generally beneficial). Our analysis derives from a critical look at standard assumptions about the epistemic states, response strategies, and strategic motivations of test takers. In conclusion, we endorse the number-right scoring rule, which discourages omissions and is robust against variability in respondent motivations, limitations in judgments of uncertainty, and item vagaries.

> Religion, politics, and formula scoring are areas where two informed people often hold opposing ideas with great assurance. (Lord, 1975, p. 7)

It is widely recognized that in multiple-choice tests, there is a probability of selecting correct answers to items about which the test taker knows nothing. A voluminous theoretical and empirical literature has developed around this guessing[1] problem (see reviews in Abu-Sayf, 1979; Diamond & Evans, 1973; Hutchinson, 1982). Most of the work focuses on the development, evaluation, and comparison of different scoring rules.

The Scoring Rules

Imagine a multiple-choice test consisting of N items with k response options each. The test taker's overt responses can be classified as Right (R), Wrong (W), or Omitted (O).[2] The simplest imaginable scoring rule, often referred to as *number right,* and here denoted $S1$, just counts the number $n(.)$ of correct

The order of authors is arbitrary. We wish to thank the following people for helpful comments on earlier drafts: Michal Beller, Gershon Ben-Shakhar, Gila Budescu, Yoav Cohen, Naomi Gafni, Baruch Nevo, David Thissen, Amos Tversky, Moshe Zeidner, and our anonymous reviewers. We also thank Ayala Cohen for providing us with room and facilities to write this article.

responses (i.e., $S1 = n(R)$). This rule is presently used by the American College Testing (ACT) and by the Graduate Record Examinations (GRE) general exams. It is both computationally and strategically simple: It is never better to omit than to answer. Omissions earn zero points, whereas a response can never earn less, while affording a positive probability of earning a point. In the terminology of decision theory, answering is a *dominant* strategy under $S1$. Nonetheless, experience and empirical studies show that, for various reasons, some examinees do not answer all items.

About 70 years ago, a different type of scoring rule was developed which incorporates a so-called correction-for-guessing feature (e.g., Holzinger, 1924; Thurstone, 1919). The basic property of these formula-scoring rules is that one's expected score is the same whether one guesses the answer to an item *at random* or whether one omits it.

One scoring rule, here denoted $S2$, levies a penalty of $1/(k-1)$ points against each incorrect answer, yielding a final score of $S2 = n(R) - n(W)/(k-1)$. This is the rule employed by the Scholastic Aptitude Test (SAT) since 1953, as well as by the GRE subject exams. $S2$ corrects for random guessing by penalizing incorrect responses, while being neutral regarding omitted items. An alternative rule, here denoted $S3$, was proposed by Traub, Hambleton, and Singh (1969). It achieves the same goal by compensating for each omission with $1/k$ points and being neutral regarding incorrect responses. Formally, $S3 = n(R) + n(O)/k$. $S3$ shares with $S2$ the property that one's expected score is the same whether one guesses at random or omits. Although in absolute terms $S3$ is higher than $S2$, they are, of course, linearly related (by the formula $S3 = [N + (k-1)\ S2]/k$). No major testing program employs $S3$. In the psychometric literature, one can find proposals for many other scoring rules, but $S1$ and $S2$ account for nearly all actual use.

The major difference between $S1$ and formula scoring, from the test taker's point of view, is that, while there is never a penalty for answering an item under $S1$, under formula scoring, an answer—should it turn out to be erroneous—earns the test taker fewer points than an omission. Hence, though ex ante neither $S2$ nor $S3$ impose a penalty on answering, ex post they impose a penalty on answering incorrectly. This presents test takers with a decision problem whenever they face an item they are not sure they can answer correctly: to guess or not to guess? The way that test takers, on the one hand, and test makers, on the other, approach this question is the subject of this article.

Our critique of scoring rules which pose this dilemma is twofold: First, we claim that it is more difficult than it seems, conceptually as well as ethically, to instruct people adequately on how to resolve this dilemma. Second, we question the psychological wisdom of scoring rules which require strategic behavior on the part of test takers, especially if the optimal application of the rules relies on subjective self-diagnoses of degrees of knowledge.

The Rationale for Formula Scoring

On the face of it, the simplicity of $S1$ would seem to make it the preferred rule for scoring multiple-choice tests. However, many regard the guessing

feature, which is intrinsic to *S*1 as problematic, on ethical or on psychometric grounds. From the test administrator's point of view, "to encourage guessing . . . is poor educational practice, since it fosters undesirable habits" (Thorndike, 1971, p. 59). From the test taker's point of view, guessing is often abhorrent, as evidenced by the many omissions that are found even under *S*1. The psychometric problem with guessing is that it interferes with what would seem to be a major goal of testing—namely, to extract the test taker's true ability from overt responses to the test. It is difficult to diagnose from a correct answer whether it reflects knowledge or luck.

*S*2 provides a partial solution to the aesthetical objections—a test taker who finds random guessing repugnant will nonetheless score the same, on average, by omitting. Moreover, under a model ubiquitous in the testing literature, and known as the *knowledge* or *random-guessing* model, *S*2 also provides an unbiased estimate of true knowledge based on test performance. According to this model, test takers either know the answer to an item, in which case they inevitably (i.e., with 100% probability) select the correct answer, or they do not, in which case they select a response alternative at random (i.e., with equal prior probabilities). Insofar as this model is untrue, however, *S*2 cannot justifiably be called correction-for-guessing (it corrects only for pure random guessing); hence it does not solve the psychometric problem that motivated it.

A Critique of the Knowledge or Random-Guessing Model

The only widely acknowledged limitation of the knowledge or random-guessing model is its failure to take into account the vast middle ground of partial knowledge that exists between full knowledge and random guessing (e.g., Davis, 1967; Lord & Novick, 1968; Nunnally, 1967). In the present section, we point out other limitations of the model and elaborate on the partial knowledge issue.

With respect to each item in a multiple-choice test, an examinee can be in one of three (subjective) states: absolute certainty, total uncertainty, or some uncertainty. In terms of the respondent's subjective probabilities, these states correspond, respectively, to being 100% sure of an answer, being equally unsure of all answers (hence, assigning a probability of $100\%/k$ to each), or having some nonuniform subjective probability distribution over the possible answers. These subjective states do not quite correspond, however, to the objective states of perfect knowledge, total ignorance, and partial knowledge, primarily because probability judgments are notoriously *miscalibrated,* a term which we now explain.

People are said to be well calibrated if they know how much they know. More formally, a judge of probabilities is calibrated if, in a (large enough) set of propositions or events to which the judge assigns a probability P%, (roughly) P% are actually true or actually occur, for any P%. It turns out, however, that people are rarely calibrated. Rather, they are biased and unreliable introspectors into their own subjective states of uncertainty (e.g., Lichtenstein, Fischhoff, & Phillips, 1982). When their subjective probabilities (or confidence ratings) are compared with their hit rates (or accuracy scores), the typical

finding is one of overconfidence: far less than P% of propositions assigned a P% subjective probability of correctness are really correct. In particular, people feel 100% certain far too often. According to estimates made by Fischhoff, Slovic, and Lichtenstein (1977) using a variety of methods (including forced choice general-knowledge tests), when people respond with 100% certainty, they are right only about 70%–80% of the time. Thus, subjective certainty does not guarantee objective accuracy, or knowledge. More generally, subjective probabilities do not correspond well enough to objective probabilities.

Another problem with miscalibration arises with regard to partial knowledge. Partial knowledge often, though by no means always, takes the form of the ability to eliminate some response options. However, the same study found that 20%–30% of options that were assigned zero probability were actually correct (Fischhoff et al., 1977). Hence, the ability to eliminate alternatives from consideration can also be overestimated. The case where a correct alternative is eliminated with great subjective certainty is called *misinformation* in the psychometric literature.

People are not only imperfectly calibrated; they also have imperfect resolution—namely, only a crude ability to distinguish between various levels of uncertainty (e.g., Lichtenstein et al., 1982). Thus, they may be unable to distinguish between a perfectly uniform subjective probability distribution (e.g., 25%, 25%, 25%, 25%) and some kind of small wobble imposed thereupon, in which one alternative or more enjoys a somewhat elevated probability at the expense of others (e.g., 31%, 23%, 23%, 23%). In other words, they may be unable to distinguish between maximal uncertainty, or total ignorance, and weak partial knowledge.

A joint consideration of miscalibration and poor resolution implies that response strategies which rely on a subjective self-diagnosis of degrees of knowledge (i.e., on experienced uncertainty) are prone to certain systematic errors—a point to which we shall return later, when considering test instructions.

The knowledge or random-guessing model is overly simplistic not only with regard to the epistemic distinctions it makes but also with regard to the process assumptions implicit in it. Even when test takers do not know the answer to a test item, it is unlikely that they choose between options "at random." The typical distribution over incorrect alternatives to an item is nonuniform, in violation of the assumption that "every wrong choice represents an unlucky *guess*" (Cronbach, 1984, p. 61, italics added). Moreover, even when test takers actually strive to guess at random, they rarely, if ever, use a random device to choose an option (though some use patterns, such as all As, or A, B, C, D in sequence). Rather, they may either choose arbitrarily, or they may attempt to fathom the presumed randomness in which correct options were assigned places (Estes, 1976). It is a robust empirical finding that people have a faulty notion of randomness and that they can neither identify nor produce random series without systematic errors (see, e.g., Bar-Hillel & Wagenaar, 1991).

A Critique of the Instructions for Formula Scoring

When formula scoring was first introduced, examinees were simply instructed not to guess. But once it was realized that sometimes guessing increases one's expected score, therefore benefiting noncompliant test takers, instructions were modified to encourage those forms of guessing. The first suggestion on how to encourage examinees to guess (Davis, 1967, p. 43) recommended the following instructions:

> Your score on this section will be based on the number of questions you answer correctly minus a fraction of the number you answer incorrectly. You should answer questions even if you are not sure your answers are correct. This is especially true if you can eliminate one or more choices as incorrect or have a hunch or feeling about which choice is correct. However, it is better to omit an item than to guess wildly among all of the choices given.

Unlike the knowledge or random-guessing model, these instructions acknowledge partial knowledge. They contain, however, an error. In terms of expected score, it is not really "better to omit an item than to guess wildly"—it is only as good. The examinee who is not given the precise value of the fraction mentioned in the instructions has no way of knowing this and might erroneously infer that the fraction is larger than $1/(k - 1)$—a necessary condition for justifying the assertion. Nowadays, the SAT uses the following adaptation of this suggestion:

> Students often ask whether they should guess when they are uncertain about the answer to a question. Your test scores will be based on the number of questions you answer correctly minus a fraction of the number of questions you answer incorrectly. Therefore, it is improbable that random or haphazard guessing will change your scores significantly. If you have some knowledge of a question, you may be able to eliminate one or more of the answer choices as wrong. It is generally to your advantage to guess which of the remaining choices is correct. Remember, however, not to spend too much time on any one question.

These instructions discard Davis's erroneous final line, but unfortunately they reduce partial knowledge to the single case where one knows enough to eliminate one or more of the contending alternatives. However, not all partial knowledge takes the form of ability to eliminate some alternatives. As we mentioned earlier, any nonuniform subjective probability distribution over response alternatives is an instance of real or perceived partial knowledge. To illustrate, consider a respondent who possesses some partial knowledge regarding an item, expressed in the following probability distribution over the four response alternatives: (40%, 20%, 20%, 20%). This is not the same as maximal uncertainty, or total ignorance, expressed as (25%, 25%, 25%, 25%). The respondent is tentatively inclined to select the first option but not confident enough to rule out any of the other options. Nonetheless, the subjective expected score from selecting the first option is higher than the expected score from either selecting an option at random or—more pertinently—from omit-

ting.[3] In this case, responding is, on average, better than omitting. Thus, ability to eliminate some alternatives is not a necessary condition for superiority of answering to omitting.

To be sure, the instructions quoted do not actually say that unless respondents can eliminate one or more alternatives they shouldn't answer, but they discourage answering inasmuch as one commits a commonplace fallacy, known in logic as "the fallacy of denying the antecedent" (e.g., Copi, 1968, p. 24). People who commit this fallacy infer (erroneously) from "If you . . . [are] able to eliminate one or more of the answer choices . . . [i]t is generally to your advantage to guess" that, if they are unable to eliminate any answer choices, it is not to their advantage to guess. Plake, Wise, and Harvey (1986, p. 20) actually committed the fallacy by telling their subjects that: "To maximize your score, if you cannot eliminate any alternatives for a particular item, you should not guess the answer but rather leave that item blank" (p. 20).

Instructions that rely on examinees' subjective feeling of certainty, as all *S*2 instructions do, require that this feeling be a trustworthy guide to choice. We mentioned above, however, that subjective probabilities are miscalibrated and exhibit systematic overconfidence. Instructions that knowingly encourage examinees to make strategic response decisions based on the fallible guide of subjective probabilities raise an ethical problem. The problem is most apparent in the context of so-called "trick questions," or "misleading items." These are items that are deliberately designed as traps for respondents with incomplete information and use distractors meant "to attract those whose knowledge and inferences are less than fully adequate" (Angoff, 1989, p. 334).

A variation on a classic trick question from the calibration literature asks, "Which is the northernmost city: New York, Denver, Rome, or Madrid?" Most respondents with some knowledge (e.g., that New York is colder than Rome, that New York is in the north of the US while Rome is in the south of Europe) erroneously eliminate Rome—the correct response.

A test that contains a preponderance of such trick items should be approached gingerly by all but the most able test takers. When faced with such items, test takers who are not of very high ability might be well advised to omit—their expected score might actually be higher thereby. (Gulliksen, 1950, saw this possibility as sufficient grounds to admonish against such items or, at the very least, against their coexistence with formula scoring.) In a study called "Does Guessing Really Help," Angoff (1989) concludes that "for high-ability students partial information may help, but for low-ability students it may hinder" (p. 334). He adds that

> we have the obligation to advise students to guess if they can *truly* eliminate one or more incorrect options. But we also have the obligation to caution them to be quite sure that the "partial information" they will use to eliminate incorrect options is *indeed* valid information. (p. 335, italics added)

He doesn't offer a concrete suggestion for how this caution could be conveyed. Indeed, unless one believes that people can distinguish between their own valid and invalid hunches, real and illusory partial knowledge, and between trick

items and ordinary items, the cautionary instructions that Angoff advocates are impossible to implement.

At this point we wish to distinguish between what we shall call *ideal* versus *real* test takers. An ideal test taker is one whose goal is to maximize expected score and whose subjective probabilities are well calibrated. A real test taker is one whose goal may differ from expected score maximization (see the following section "Beyond Expected Scores") and whose subjective probabilities are not necessarily well calibrated. It is possible to be a rational test taker without being an ideal test taker. In decision theory, *rationality* is defined in terms of the matching of means to goals, not in terms of the goals themselves. There is nothing irrational, therefore, about having a test goal that differs from maximizing expected score, but miscalibration is irrational.

In the psychometric literature, the dominant model of the test taker is that of an ideal test taker. Nonetheless, hedges such as the word *generally* in the SAT instructions may hint at some unease with this model. For an ideal test taker, this hedge is superfluous, since the statement it precedes is unqualifiedly true. It is only required for misinformed test takers—but misinformation is a state of uncertainty in which well-calibrated test takers can never find themselves (i.e., it cannot happen that distractors which enjoy a subjective probability lower than $1/k$ will be systematically likely to be correct). No *S*2 instructions tackle the possibility of misinformation head on.

Beyond Expected Scores

Hitherto, it was assumed that the test taker is an expected score maximizer. In what follows, we give several scenarios that undermine the exclusivity of this criterion. We show that a test taker could have legitimate preferences between guessing and omission even when the two have equal expected values—or could even prefer the strategy with the lower expected value.

a. Risk preferences: Recall that, whereas the score for an omitted item is a constant (0 under *S*2, and $1/k$ under *S*3), the score for guessed items is a random variable (it can be either 1 or $-1/(k-1)$ under *S*2, and either 1 or 0 under *S*3). In other words, whereas there is no variability in the scores for omitted items, there is some in the score for guessed items. One reason people may prefer one response strategy over the other is that they may have risk preferences. For example, risk aversion (the preference for a sure thing over a gamble in which the expected value is at least as large as the sure thing) is a common phenomenon (e.g., Kahneman & Tversky, 1979) that entails a preference for omitting over guessing—even at some loss in expected value. The tendency to guess rather than omit has also been found to be somewhat correlated with gender (females are less likely to guess) and cultural background (minorities are less likely to guess; see Ben-Shakhar & Sinai, 1991; Gafni & Melamed, 1990; Grandy, 1987), in a manner suggestive of systematic individual differences in risk taking behavior.

b. Passing a predetermined cutoff level: Whereas the expected score for responding is never lower than for omitting, the actual score for a given choice of answer can be lower—or higher—than for omitting (in fact, for a single item,

it is never equal). Imagine applicants who are taking a licensing test under $S2$ with a known and predetermined passing score, say 80%. If the applicants believe that after having answered a certain number of questions they have already accumulated enough points to pass (say 90%), they have little to gain by attempting to increase their score further and could jeopardize it by guessing wrongly. This is a valid argument for omitting—not only items they know nothing about, but even items they have some knowledge about. Conversely, other examinees who believe that the items they know for sure place them far short of the passing score (say at 70%) are better off guessing the rest, even at random, since in terms of their goal, if not in terms of expected score, they have more to gain than to lose thereby.

c. All-or-none payoff: An applicant is vying for a prize which will be awarded only for a perfect score (i.e., N correct answers). Clearly, the only strategy with a nonzero probability of achieving a perfect score is to attempt all items, regardless of the scoring rule. In this case, guessing would be superior to omission for any scoring rule.

d. Competitiveness: Imagine a candidate who wants to select that response strategy of two (guess all or omit all) which enjoys the higher probability of outscoring the other. It turns out (see appendix) that, depending on the values of N and of k, one of these strategies may outperform the other, although their expected value is the same. For example, if $N = 3$ and $k = 4$, the probability that guessing will outscore omitting is .58. On the other hand, if $N = 5$ and $k = 4$, the probability that omitting will outperform guessing is .63.

In these scenarios, instructions tailored exclusively to an expected score criterion would be inappropriate, but it is difficult to give formula scoring instructions that take account of them all. Under $S1$, on the other hand, the recommendation to guess can be made no matter what.

A Comparison of the Scoring Rules

Scoring rules can be compared on three dimensions: strategic, psychological, and psychometric. *Strategic* refers to the way a rational decision maker (i.e., an ideal test taker) ought to respond to the rule; *psychological* refers to the way real test takers respond to the rule; *psychometric* refers to the reliability and validity of tests scored by these rules. In this section, we will compare the scoring rules on these three dimensions. First, we compare the two formula scoring rules, $S2$ and $S3$.

Ideal test takers operating under formula scoring should answer any item about which they have complete or partial knowledge. Recall that, in the general case of complete or partial knowledge, not all alternatives are assigned the same subjective probability. There is necessarily an alternative the subjective probability of which exceeds $1/k$, where k is the number of alternatives. Under $S3$, the score for an omission is $1/k$, whereas selecting the most probable alternative has an expected score in excess of $1/k$; hence answering is superior to omitting. Under $S2$, the score for omission is 0, whereas the expected score for answering exceeds 0; so again, answering is superior to omitting. On the other hand, if test takers are going to guess at random, then their expected

score is exactly the same whether they choose to guess or to omit, and they should be indifferent between the two. Insofar as this condition for omitting is identical for both formula scores, they are strategically equivalent.

In spite of their strategic equivalence, there is some empirical evidence that *S*3 yields higher reliabilities (e.g., Traub & Hambleton, 1972) and validities (Sax & Collet, 1968) than *S*2. In addition, test takers seem to prefer *S*3 to *S*2 (Waters & Waters, 1971). Why would two scoring formulas that are strategically equivalent and perfectly correlated result in different psychometric properties? The answer lies in the psychological dimension.

In economic theory, an *opportunity cost* is the name given to failure to realize a possible gain. For example, if one purchased an item for $\$X$ when one could just as easily have purchased it for $\$X$-$\Delta$, then Δ is the opportunity cost of this transaction. "The first lesson of economic theory is that all costs are (in some sense) opportunity costs. Therefore opportunity costs *should* be treated as equivalent to out-of-pocket costs" (Thaler, 1980, p. 44). In contrast, real people experience opportunity costs (e.g., failure to win a bonus) quite differently from out-of-pocket costs (e.g., paying a penalty), typically underweighting the former relative to the latter. For example, people are less reluctant to charge a purchase when the difference between cash and credit prices is called a "cash discount" than when it is called a "credit surcharge," a fact well appreciated by the credit card industry (Thaler, 1980). Similarly, the difference in tax rates between people with children and childless people seems much more acceptable when labeled a "child exemption" rather than a "childless premium," although clearly these are equivalent (Schelling, 1981). Generalizing to the present context, it is likely that real test takers do not consider a foregone bonus for omission as threatening as a penalty for error and are therefore more willing to guess when an error will exact the former (opportunity) cost than when it will exact the latter (out-of-pocket) cost.

We turn now to a comparison of formula scoring with *S*1. If an ideal test taker omits no items under formula scoring, there is no reason to omit them under *S*1 either. Suppose, however, that some items were omitted under formula scoring. For an ideal test taker, the omitted items would be only those that would have been answered at random under *S*1. In that case, all three scores would be "unbiased estimators of the same quantity" (Lord, 1975, p. 9), and the only purely psychometric grounds for choosing between them would be their variances. Since variance due to random guessing reduces reliability, *S*3, with the least variance, would be the rule of choice,[4] and *S*1 (with the largest error component due to the largest guessing rates) would be least preferred.

Formula scoring was not introduced merely to reduce error variance, and it is psychologically implausible that this is all it does. It is equally unlikely that the three rules elicit the same kind of response behavior with respect to partially known items—as ideally they should. The psychometric comparisons of formula scoring versus number right have focused on the question of whether test takers are better off under *S*1 or under formula scoring—primarily *S*2.

There are two paradigms for comparing the two scoring rules—between and within. In the within paradigm, examinees first do a test under *S*2 instructions

and then are encouraged to return to omitted items and try to answer them. Scores have usually increased, indicating that examinees were not merely guessing randomly (e.g., Cross & Frary, 1977; Slakter, 1968). The between paradigm, in contrast, has not yielded evidence that items omitted under formula scoring would have been answered better than chance (Angoff & Shrader, 1984). However, Albanese (1986) has challenged the generalizability of Angoff and Schrader's results on methodological grounds.

There have been some simulation studies carried out to examine the impact of formula scoring on an individual's score (Albanese, 1988; Frary, 1980). The results do not, of course, show anything that could not have been deduced analytically (e.g., that under valid partial information answering is beneficial, while under misinformation it is not). They are useful, however, in giving quantitative estimates of the impact of formula scoring on different examinees of differing abilities undertaking items of varying degrees of difficulty. Under some conditions, the impact could make a considerable difference.

Additional evidence for different test taking behavior under $S1$ and $S2$ comes from looking at response time. Ben-Simon (1992) recently administered four different types of tests (general knowledge, figures, reasoning, quantitative ability) under instructions for either $S1$ or $S2$ to eight groups (test type $\times$ scoring rule type) of about 100 respondents each. The average time required to complete the test was almost 10% higher under formula scoring than under number-right scoring. Angoff and Schrader (1984) report that, in speeded tests, trailing omits (i.e., items following the last answered item, which were probably never even reached) were about 25%–30% more numerous under $S2$ than under $S1$. These results suggest that the very need to decide whether to answer an item or omit it costs time (Albanese, 1986, 1988).

Conclusions

Guessing is bad for test makers, not necessarily for test takers. Formula scoring was initially developed to discourage guessing. For the ideal test taker, however, formula scoring merely obviates guessing—and only random guessing at that. To really discourage guessing, the penalty for errors should exceed $1/(k-1)$. Indeed, for any level of certainty, there exists a penalty for error that makes it not worth the test taker's while to answer with less than that certainty. For $S2$, that level is $1/k$—which is the degree of certainty associated with random guessing. Therefore, any guessing but random guessing is, on average, worthwhile under $S2$. We titled this article "To Guess or Not to Guess?" For an ideal test taker, the answer to this question is clear—guess![5]

If the question "To guess or not to guess" must be faced by test takers, test makers must face the question "To correct or not to correct for guessing?" We have argued that, if the goal of correction for guessing is to discourage guessing, the goal is prejudicial and the means less than appropriate. But if even under $S2$ guessing should be encouraged, not discouraged, then its sole advantage to test makers is the variance saved on omitted items. Were all test takers ideal, $S2$—with appropriate instructions!—would be preferable to $S1$.

Test takers, however, are generally miscalibrated and occasionally not even

expected score maximizers. The first factor means that test takers sometimes omit when they shouldn't (i.e., when their expected score from answering would have been higher) and sometimes answer when they shouldn't (i.e., when they are misinformed). But, short of giving test takers a crash course on formula scoring, it is nearly impossible to give recommendations that will be fair and beneficial to all.

Some researchers have taken a position that test instructions should provide a full and accurate description of the scoring rule but stop short of actual strategic advice (see Abu-Sayf, 1979, for a review). *S*2 instructions seldom give the scoring formula explicitly (but see Abu-Sayf, 1977). They may mention the loss of "a fraction of the number [of items] you answer incorrectly," but they fail to give that fraction. This is vital information. For whether the fraction is smaller than, larger than, or equal to $1/(k - 1)$ determines whether random guessing has an expected value higher than, lower than, or equal to omitting. Unless this is known, even perfectly rational test takers cannot figure out their optimal strategy, and they are totally reliant on the accuracy, adequacy, and fairness of the instructions.

On the other hand, evidence shows that test takers cannot always be counted on to draw correct strategic implications from scoring rules. A notable case in point are test takers who omit under *S*1 (e.g., in 1984, shortly after changing their scoring rule from *S*2 to *S*1, only 44% of GRE examinees answered all questions, and as many as 5%—about 3,000 examinees—failed to answer 20 items or more; Grandy, 1987).

Failing to advise test takers may be doing them a disservice. Advising them badly is an even greater disservice. *S*1 affords a way out of this quandary. It is the only scoring rule that allows one to make simple, straightforward, unqualified recommendations regarding response strategy that are robust to all conceivable differences in motivations and abilities, as well as impervious to the cognitive limitations under which test takers labor and to the quirks of test items. Considering that even test makers have occasionally failed to deal satisfactorily with the strategic dilemmas raised by formula scoring, the opportunity to avoid this dilemma altogether becomes an increasingly attractive one.

We believe that, if test administrators were to take an unambivalent stance upholding answering regardless of epistemic state, the test taking public could, and soon would, be educated in that spirit, and reluctance to guess would diminish if not disappear. "[I]f a culture were consistent in rewarding the examinee for answering all items, few would fail to respond in that way" (Thorndike, 1971, p. 61). In this respect, it is enlightening to note the Israeli experience. The Psychometric Entrance Test (PET), used for selecting among applicants to Israeli universities, employs *S*1. In 1984, its first year, it reported nonresponse rates of 30%–40% (depending on the subtest), which dropped over the years to 5%–20%, apparently as a simple consequence of the rising test sophistication of the applicant population (Alalouf & Sadan, 1990). This happened without any special measures directed explicitly toward this goal. Combined with an explicit urging to answer, answer, answer, these proportions might well drop even further.

Additionally, *S*1 has the virtue, not shared by formula scoring, that noncompliance is easily and unambiguously detectable. Hence, follow-up studies of compliance rates, as well as measures such as sending noncompliers back to their seats with an encouragement to answer the omitted items, become feasible. Even if noncompliance can never be totally eliminated, we believe that it is a lesser evil than the ones accompanying the strategic pitfalls of formula scoring. Test makers owe it to test takers to give instructions that are both irreproachable and useful—which for *S*2 is practically impossible to do. Test takers, on the other hand, owe it only to themselves to comply with those instructions.

In our discussion, we paid little attention to the moral and educational aspects of guessing. *S*1 in effect levies a penalty against those reluctant to guess. Drawing the line between the kind of guessing that should be encouraged (e.g., mining partial knowledge) and the kind that should, perhaps, not be (e.g., capitalizing on chance) is very difficult. Moreover, test takers who choose to disregard this distinction cannot be prevented from doing so. Encouraging all test takers to answer all items—including those unread—at least removes the advantage that the shrewd, bold, or entrepreneurial have over the shy, inhibited, or cautious.

Notes

[1]To *guess* in a multiple-choice test can have either the everyday nontechnical meaning of to select an answer without being sure of its correctness or the technical meaning of to select one of a set of answers with equal probabilities. The latter has often been called *wild guessing* or *blind guessing.* Here we shall also call it *random guessing* or *pure guessing* to distinguish it, where necessary, from the noncommittal guessing.

[2]We ignore a distinction sometimes made between pure omissions and items not reached, since they both affect the scoring in precisely the same way.

[3]Since the test taker believes the first answer is most likely to be correct, it will be the one chosen. The test taker's expected score would then be .2 (i.e., there is a .4 subjective probability for answering correctly and scoring a point and a .6 probability for erring and incurring the penalty of $-1/k$; altogether, $.4 \times 1 - .6 \times \frac{1}{3} = .2$) under *S*2, and .4 under *S*3. The expected score from random guessing is 0 (i.e., $.25 \times 1 - .75 \times \frac{1}{3}$) under *S*2, and .25 under *S*3. Likewise, the score for omitting is 0 under *S*2, and .25 under *S*3. Under *S*2, $.2 > 0$, and under *S*3, $.4 > .25$.

[4]The variance of the guessing component under *S*2 is $1/(k - 1)$, which is greater than $(k - 1)/k^2$, its counterpart under *S*3, for any k. Hence, in the sense that the variance of a gamble is often taken as a measure of its riskiness (e.g., Pollatsek & Tversky, 1970), there is more risk associated with *S*3 than with *S*2, which could also contribute to a preference of risk-averse people for *S*3 over *S*2.

[5]We have occasionally been asked about the applicability of our conclusions to speeded testing or to very difficult tests. Such tests present test takers with a severe time allocation problem. The solution to this problem is orthogonal to the scoring rule issue. If—and only if—one decides to omit an item under *S*2, one should blindly guess an answer under *S*1. Note that this does not require that one read either the item or the answers. The time required for truly blind guessing is negligible—but it is essential to instruct examinees appropriately to that effect.

APPENDIX

Consider an examinee, who is in a state of total uncertainty regarding n items, taking a multiple-choice test under formula scoring S_3. Which strategy is more likely to outscore the other: choosing to omit all n items or guessing them all at random? Under S_3, omission guarantees a score of $O = n/k$. The score for guessing, G, is a random variable. Our question, then, is when is $P(G > O) > 1/2$?

The probability of guessing correctly exactly m of the n items is given by the binomial $B(m|n, 1/k) = \binom{n}{m}\left(\frac{1}{k}\right)^m\left(\frac{k-1}{k}\right)^{n-m}$. Thus, the probability that $G \leq O$ is obtained by summing all the terms of the binomial distribution from $m = 0$ to $m = [O]$ ($[O]$ is the integer part of O). Similarly, the probability that $G > O$ is obtained by summing all the remaining terms of the same distribution, from $m = [O] + 1$ to $m = n$. If $k = 2$ (as in true/false tests), G is distributed symmetrically around O, so that $P(G > O) = P(O > G)$.

For $k > 2$, the result depends on the binomial parameters, as follows.

1. If n is an integer multiple of k, there is always a large enough probability that the two actions yield equal scores. Consequently, there is hardly an advantage to one of the actions over the other, though omitting enjoys a consistent albeit slight advantage.
2. If $n(\text{mod } k) = 1$, it pays to guess—namely, $P(G < O) < 0.5 < P(O < G)$. In other words, in this case random guessing has a better than even chance of outscoring riskless omission.
3. In all other cases, it is better to omit.

As n increases or, for any fixed value of n, as k decreases, the ratio of $P(G > O)$ to $P(O > G)$ converges toward unity. Hence, the sharpest differences between random guessing and omitting will appear for multiple-choice items with many response options and for highly knowledgeable examinees (those who need guess only a small number of items).

References

Abu-Sayf, F. K. (1977). A new formula score. *Educational and Psychological Measurement, 37,* 853–862.

Abu-Sayf, F. K. (1979). The scoring of multiple-choice tests: A closer look. *Educational Technology, 19,* 5–15.

Alalouf, A., & Sadan, Y. (1990). *Changes in PET score distributions across years* (Report No. 123). Jerusalem, Israel: National Institute for Testing and Evaluation.

Albanese, M. A. (1986). The correction for guessing: A further analysis of Angoff and Schrader. *Journal of Educational Measurement, 23,* 225–235.

Albanese, M. A. (1988). The projected impact of the correction for guessing on individual scores. *Journal of Educational Measurement, 25,* 149–157.

Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement, 26,* 323–336.

Angoff, W. H., & Schrader, B. W. (1984). A study of hypotheses basic to the use of rights and formula scores. *Journal of Educational Measurement, 21,* 1–17.

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

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28,* 23–35.

Ben-Simon, A. (1992). *Psychometric and cognitive aspects of partial knowledge in solving multiple-choice tests.* Unpublished doctoral dissertation, The Hebrew University, Jerusalem.

Copi, I. M. (1968). *Introduction to logic.* London: Macmillan.

Cronbach, L. J. (1984). *Essentials of psychological testing.* New York: Harper & Row.

Cross, L. H., & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *Journal of Educational Measurement, 14,* 313–321.

Davis, F. B. (1967). A note on the correction for chance success. *Journal of Educational Measurement, 3,* 43–47.

Diamond, J. J., & Evans, W. (1973). The correction for guessing. *Review of Educational Research, 43,* 181–191.

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

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564.

Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement, 4,* 79–90.

Gafni, N., & Melamed, E. (1990). *Differential tendencies to guess as a function of gender and lingual-cultural reference group* (Report No. 148). Jerusalem, Israel: National Institute for Testing and Evaluation.

Grandy, J. (1987). Characteristics of examinees who leave questions unanswered on the GRE general test under right-only scoring (Research Report No. 87-38). Princeton, NJ: Educational Testing Service.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Holzinger, K. J. (1924). On scoring multiple-response tests. *Journal of Educational Measurement, 15,* 445–447.

Hutchinson, T. P. (1982). Some theories of performance in multiple-choice tests and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology, 35,* 71–89.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12,* 7–11.

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

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Plake, B. S., Wise, S. L., & Harvey, A. L. (1986, April). *Investigation of test-taking behavior: Complementary research approaches.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Pollatsek, A., & Tversky, A. (1970). A theory of risk. *Journal of Mathematical Psychology, 7,* 540–553.

Sax, G., & Collet, L. (1968). The effects of different instructions and guessing formulas on reliability and validity. *Educational and Psychological Measurement, 28,* 1127–1136.

Schelling, T. C. (1981). Economic reasoning and the ethics of policy. *Public Interest, 63,* 37–61.

Slakter, M. (1968). The penalty for not guessing. *Journal of Educational Measurement, 5,* 141–144.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization, 1,* 39–60.

Thorndike, R. L. (Ed.). (1971). *Educational measurement.* Washington, DC: American Council on Education.

Thurstone, L. L. (1919). A method for scoring tests. *Psychological Bulletin, 16,* 235–240.

Traub, R. E., & Hambleton, R. K. (1972). The effects of scoring instructions and degree of speededness on the validity and reliability of multiple-choice tests. *Educational and Psychological Measurement, 32,* 737–758.

Traub, R. E., Hambleton, R. K., & Singh, D. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *Educational and Psychological Measurement, 29,* 847–862.

Waters, C. W., & Waters, L. K. (1971). Validity and likeability ratings for three scoring instructions for a multiple-choice vocabulary test. *Educational and Psychological Measurement, 31,* 935–938.

Authors

DAVID BUDESCU is Associate Professor, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign, IL 61820. *Degrees:* BA, University of Haifa, Israel; MA, PhD, University of North Carolina. *Specializations:* quantitative psychology, and judgment and decision making.

MAYA BAR-HILLEL is Associate Professor, Department of Psychology, The Hebrew University, Jerusalem, Israel 91905. *Degrees:* BA, MA, PhD, The Hebrew University, Israel. *Specializations:* probabilistic reasoning, rationality, and distributive justice.