

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 episodic 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 guessing¹ 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 S_1 , just counts the number $n(\cdot)$ 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 Zeldner, and our anonymous reviewers. We also thank Ayalia 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 $S1$ 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" (Thurstone, 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 $S1$. 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.

$S2$ provides a partial solution to the aesthetic 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, $S2$ 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, $S2$ 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

