Zapechelnyuk, A. (2015) An axiomatization of multiple-choice test scoring. *Economics Letters*, 132, pp. 24-27.

Copyright © 2015 Elsevier B.V.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

Content must not be changed in any way or reproduced in any format or medium without the formal permission of the copyright holder(s)

When referring to this work, full bibliographic details must be given

http://eprints.gla.ac.uk/104929/

Deposited on: 4 June 2015

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

AN AXIOMATIZATION OF MULTIPLE-CHOICE TEST SCORING

Andriy Zapechelnyuk

University of Glasgow

Abstract. This note axiomatically justifies a simple scoring rule for multiple-choice tests. The rule permits choosing any number, k, of available options and grants $1/k$-th of the maximum score if one of the chosen options is correct, and zero otherwise. This rule satisfies a few desirable properties: simplicity of implementation, non-negative scores, discouragement of random guessing, and rewards for partial answers. This is a novel rule that has not been discussed or empirically tested in the literature.

JEL Classification: C44, A2, I20

Keywords: multiple-choice test, scoring rules, axiomatic approach.

1. Introduction. Multiple-choice questions are routinely used in examinations. They are simple to implement and score, and do not have apparent disadvantages relative to essay questions (Akeroyd 1982, Bennett, Rock, and Wang 1991, Bridgeman 1991, Walstad and Becker 1994, Brown 2001).

A multiple-choice question seeks a single correct answer from a list of options. Multiple-choice questions are almost universally evaluated by the *number right* scoring rule that grants the unit score if a single correct option is chosen and zero otherwise. This method suffers from recognized drawbacks: it encourages guessing and does not permit expressing partial knowledge. From a test-maker's point of view, this is undesirable, as it interferes with inference of true knowledge of a test-taker from her response to the test. A correct answer may equally signify knowledge and luck. From a test-taker's point of view, this is also undesirable. A risk averse test-taker who is hesitating between a few answers is forced to gamble to grab her chance and cannot opt for a lower, but more certain score.

The problem of guessing is traditionally addressed by penalizing wrong answers with negative scores, called *formula scoring* (e.g., Holzinger 1924). This approach is implemented, for example, in the SAT and GRE Subject tests. Interestingly, formula scoring does not really solve the problem: if a risk-neutral test-taker can eliminate some options but hesitates among the remaining ones, she strictly prefers to make a

Date: March 14, 2015.

Zapechelnyuk: Adam Smith Business School, University of Glasgow, University Avenue, Glasgow G12 8QQ, UK. *Phone:* +44(0)141-330-5573. *E-mail:* azapech@gmail.com

Acknowledgements: I would like to thank David Budescu, Oscar Volij, and Ro'i Zultan for helpful comments.

guess (Budescu and Bar-Hillel 1993, Bar-Hillel, Budescu, and Attali 2005). Negative scores per se have also been criticized for contributing to high omission rates and discrimination against risk-averse and loss-averse test-takers (Ben-Simon, Budescu, and Nevo 1997, Burton 2005, Delgado 2007, Budescu and Bo 2014).[1]

Anther well-known scoring method that discourage guessing and elicit partial knowledge is *subset selection scoring* (Dressel and Schmidt 1953),[2] which allows a test-taker to choose a subset of options and grant score 1 for the correct option and $-\frac{1}{n-1}$ for each incorrect option in the chosen set. The literature also studies complex scoring methods that elicit test-takers' ordinal ranking, confidence, or probability distribution over available options (Bernardo 1998, Alnabhan 2002, Swartz 2006, Ng and Chan 2009). Though such scoring rules are advantageous in theory, the evidence suggests that they might not be advantageous in practice (Budescu and Bar-Hillel 1993, Bar-Hillel, Budescu, and Attali 2005, Espinosa and Gardeazabal 2013, Budescu and Bo 2014). The problem is a distortion between the inference from responses and the true knowledge caused by test-takers' strategic considerations. With complex scoring, test-takers' responses depend not only on their knowledge, but also on specifics of the scoring rule and personal characteristics (risk attitude, loss aversion, etc.).

To sum up, there are a few desirable properties of multiple-choice scoring:

(a) simplicity;
(b) non-negative scores;
(c) discouragement of guessing;
(d) rewards for partial answers.

This note axiomatically derives a scoring rule that satisfies the above properties. The rule permits to select any number k out of n available options and grants $1/k$-th of the maximum score if one of the chosen options is correct, and zero otherwise. This rule is uniquely determined by a simple requirement. A risk-averse test-taker who is indifferent between a few options should prefer to choose all of them, rather than choosing either of them (and "prefer" replaced by "indifferent" for a risk-neutral test-taker).

To the best of our knowledge, this rule has not been discussed or empirically tested in the literature. It is a variant of subset selection scoring mentioned above, however, it assigns scores to selected subsets differently, and therefore has different properties. Most notably, subset selection scoring discourages guessing "too much" and penalises wrong answers harsher than our rule (see more details in Section 3). Thus, our scoring rule, at least hypothetically, evokes less distortion of responses due to strategic considerations of test-takers.

Frandsen and Schwartzbach (2006) propose a different axiomatization of multiple-choice scoring. The two defining axioms of Frandsen and Schwartzbach (2006) are

[1] For an alternative opinion see Espinosa and Gardeazabal (2010).

[2] Equivalent variants are *elimination scoring* (Coombs, Miholland, and Womer 1956, Bradbard and Green 1986) and *liberal scoring* (Bush 2001, Bradbard, Parker, and Stone 2004, Jennings and Bush 2006).

invariance under decomposition (if a question is decomposable into two simpler questions, then the score of the complex question is the sum of the scores of the simpler ones) and *zero sum* (the expected score of random guessing is zero). As a result, a choice of k out of n available options gives score $\ln\left(\frac{n}{k}\right)$ if it contains the correct answer and $-\frac{k}{n-k}\ln\left(\frac{n}{k}\right)$ otherwise. This scoring rule has a very nice interpretation from the information-theoretical perspective. Yet it permits negative scores and qualitatively compares to our rule in the same way as the subset selection scoring (see more details in Section 3).

2. The scoring rule. A test-taker is permitted to choose any number of options out of $n \geq 2$ available; only one option is correct.

A scoring rule assigns a numerical value $f_z(k)$ to a choice of k out of n options, where $z \in \{0,1\}$ indicates whether the chosen set contains the correct answer ($z=1$) or not ($z=0$). The number of options, n, is fixed and omitted from notation.

We assume that scoring functions satisfy two primitive properties. First, we normalize the scores to be in $[0,1]$ and assume that the maximum is achieved by choosing the single correct option, while the minimum is achieved by choosing $n-1$ incorrect options:

$$f_1(1) = 1 \text{ and } f_0(n-1) = 0. \tag{1}$$

Second, two equally uninformative responses, selecting all options and omitting the question, should be scored equally:

$$f_1(n) = f_0(0). \tag{2}$$

Denote by $\mathcal{F}$ the set of scoring functions that satisfy the above properties.

We now describe the choice of a test-taker. Denote by $N = \{1,\dots,n\}$ the set of available options. Let $p = (p_a)_{a\in N}$ be a probability vector. The test-taker believes that each answer a is correct with probability p_a.

The test-taker has to chooses a subset $A \subset N$ (possibly, $A = N$ or $A = \varnothing$). The test-taker is risk-averse (or risk-neutral) and evaluates a choice set $A \subset N$ under a probability vector p according to the expected utility:

$$U(A,p) = p_A u(f_1(|A|)) + (1-p_A)u(f_0(|A|)),$$

where $p_A = \sum_{a\in A} p_a$ and $u : [0,1] \to \mathbb{R}$ is a utility function. We assume that u is continuous and weakly concave, and normalize

$$u(0) = 0 \text{ and } u(1) = 1. \tag{3}$$

We say that the test-taker *prefers A to B (strictly prefers, indifferent) under probability vector p* and use notation $A \succsim_p (\succ_p, \sim_p) B$ if

$$U(A,p) \geq (>,=) U(B,p).$$

We now impose a requirement (axiom) on the test-taker's choice that formalizes the idea that test-takers should be discouraged from random guessing: "If you don't know which answer to choose, then choose both." A test-taker should prefer to choose all options about which she is indifferent, rather than choosing any single one.

Axiom 1. *If for some probability vector p, some $A \subset N$, and some $a \in A$,*

$$a \sim_p b \quad \text{for all } b \in A,$$

then

$$\begin{aligned} A &\succsim_p a \quad \text{under risk-averse preferences,} \\ A &\sim_p a \quad \text{under risk-neutral preferences.} \end{aligned}$$

Essentially, when all options in A are equally likely to be correct, Axiom 1 requires that choosing A yields the same expected score as the lottery associated with choosing any single option $a \in A$.

The consequent $A \sim_p a$ for a risk-neutral individual makes the axiom tight—a risk-loving test-taker would actually prefer random guessing to choosing set A.

Axiom 1 pin down a unique scoring rule in $\mathcal{F}$.

Theorem 1. *The unique scoring rule in $\mathcal{F}$ that satisfies Axiom 1 is given by*

$$f_1(k) = \frac{1}{k} \quad \text{and} \quad f_0(k) = 0$$

for every $k \in \{1, \ldots, n\}$, and $f_0(0) = \frac{1}{n}$.

This scoring rule is simple, even relative to subset selection and elimination scoring, and admits only nonnegative scores by design. It also discourages guessing, as whenever a test-taker is indifferent between two disjoint sets A and B, she prefers to choose both of them:

$$A, B \subset N \text{ are disjoint and } A \sim_p B \quad \Longrightarrow \quad A \cup B \succsim_p A.$$

Finally, this scoring rule rewards partial answers: a test-taker who can narrow down her choice to a subset A of options but unsure about choosing within A gets a partial credit for choosing the whole A.

3. Related scoring rules. To the best of our knowledge, the scoring rule in Theorem 1, as well as its re-normalized version $\tilde{f}$ that gives zero score to omission,[3]

$$\tilde{f}_1(k) = \frac{n-k}{(n-1)k} \quad \text{and} \quad \tilde{f}_0(k) = \begin{cases} -\frac{1}{n-1}, & k \geq 1, \\ 0, & k = 0, \end{cases}$$

has not been previously discussed nor empirically tested.

A closely related scoring rule is *subset selection scoring*. It grants score 1 for the correct option and $-\frac{1}{n-1}$ for each incorrect option in the chosen set. For a choice of k out of n options it is:

$$g_1(k) = 1 - \frac{k-1}{n-1} \quad \text{and} \quad g_0(k) = -\frac{k}{n-1}.$$

Frandsen and Schwartzbach (2006) use the axiomatic approach to derive the *logarithmic scoring rule* as the only one that satisfies the axioms of *invariance under decomposition* (if a question is decomposable into two simpler questions, then the

[3] Set the maximum score to 1 and the omission score ($A = \varnothing$) to 0.

score of the complex question is the sum of the scores of the simple ones) and *zero sum* (the expected score of random guessing is zero):

$$h_1(k) = \ln\left(\frac{n}{k}\right) \quad \text{and} \quad h_0(k) = -\frac{k}{n-k}\ln\left(\frac{n}{k}\right).$$

The above two rules reward correct answers more generously, but also penalize incorrect answers more severely, as compared to our rule. Particularly, for each number of chosen options k, scores assigned by the subset selection scoring rule are k times as large as scores assigned by our rule, $g_z(k) = k\tilde{f}_z(k)$.

There are two important consequences of this difference. First, scoring rules g and h discourage random guessing "too much", and hence violate our Axiom 1. For example, consider a multiple-choice question with the set of options $N = \{a_1, a_2, a_3, a_4, a_5\}$, and assume that a risk-neutral test-taker has beliefs $p = (\frac{1}{2}, \frac{1}{2}, 0, 0, 0)$. Axiom 1 demands that $\{a_1\} \sim_p \{a_1, a_2\}$. But under both g and h, $\{a_1\} \prec_p \{a_1, a_2\}$.

Consider a more drastic example. Let $p = (\frac{2}{3}, \frac{1}{3}, 0, 0, 0)$, that is, the test-taker believes that option a_1 is twice as likely to be correct as option a_2, and the rest of options are surely incorrect. One may expect that a risk-neutral (or close to risk-neutral) test-taker would prefer the likely option a_1 to the set $\{a_1, a_2\}$. This is indeed the case under our scoring rule. However, $\{a_1\} \prec_p \{a_1, a_2\}$ under both g and h for every risk-averse or risk-neutral test-taker.

Second, as compared to our rule, scoring rules g and h generate a higher variance of lotteries, and thus evoke more distortion between the inference from responses and the true knowledge caused by strategic considerations of risk-averse and loss-averse test-takers (Budescu and Bar-Hillel 1993, Budescu and Bo 2014).

Finally, scoring rules g and h admit negative values. There is some evidence suggesting that negative scoring is undesirable, particularly, due to discrimination against loss-averse test-takers (e.g., Delgado 2007, Budescu and Bo 2014). The re-normalization of the score range to $[0, 1]$ interval uncovers another potential problem: these rules are too lenient on test-takers who know nothing. For an uninformative answer or omission, the normalized subset selection rule gives score $1/2$ irrespective of n, while the normalized logarithmic scoring rule yields the score, for example, about $\frac{1}{3}$ for $n = 6$ and about $\frac{1}{4}$ for $n = 18$. In contrast, for an uninformative answer or omission, our scoring rule yields score $\frac{1}{n}$ for every n.

4. Proof of Theorem 1. We prove that the scoring rule stated in Theorem 1 is the only one that satisfies Axiom 1 for an individual with risk-neutral preferences, $u(x) = x$. Then we show that this scoring rule also satisfies Axiom 1 for any risk-averse individual.

The expected utility of a risk-neutral test-taker from choosing set A is equal to the expected score:

$$U(A, p) = p_A f_1(|A|) + (1 - p_A) f_0(|A|).$$

For every $a \in N$ we have $U(a, p) = p_a f_1(1) + (1 - p_a) f_0(1)$, hence, $a \sim_p b$ if and only if $p_a = p_b$. Consider a probability distribution p that is uniform on some subset

A, $p_a = \bar{p}$ for all $a \in A$. Denote $k = |A|$. Then Axiom 1 implies that for every $k = 2, 3, \dots, n-1$ and every $\bar{p} \in [0, \frac{1}{k}]$

$$\bar{p} f_1(1) + (1-\bar{p}) f_0(1) = k\bar{p} f_1(k) + (1 - k\bar{p}) f_0(k),$$

or equivalently,

$$\bar{p}\Big(f_1(1) - f_0(1) - k(f_1(k) - f_0(k))\Big) + f_0(1) - f_0(k) = 0.$$

Since the above has to hold for all $\bar{p} \in [0, \frac{1}{k}]$, we have

$$f_1(1) - f_0(1) - k(f_1(k) - f_0(k)) = 0, \quad k \in \{2, \dots, n-1\}. \tag{4}$$

and

$$f_0(1) - f_0(k) = 0, \quad k \in \{2, \dots, n-1\}. \tag{5}$$

Recall that $f_0(n-1) = 0$ by (1), hence (5) implies

$$f_0(k) = 0, \quad k \in \{1, \dots, n-1\}.$$

Also recall that $f_1(1) = 1$ by (1), hence (4) becomes $1 = kf_1(k)$, and consequently,

$$f_1(k) = \frac{1}{k}, \quad k \in \{2, \dots, n-1\}.$$

Finally, (2) implies $f_0(0) = f_1(n) = \frac{1}{n}$.

We now verify that Axiom 1 is satisfied for a risk-averse test-taker. Let f be as defined above. Let A be a set such that the test-taker is indifferent between any of its options: for every $a, b \in A$,

$$p_a u(f_1(1)) + (1-p_a) u(f_0(1)) = p_b u(f_1(1)) + (1-p_b) u(f_0(1)).$$

Since $f_1(1) = 1$ and $f_0(1) = 0$ and we have $u(0) = 0$ and $u(1) = 1$ by (3), the above holds if and only if $p_a = p_b$. Thus, we have p_a are the same for all $a \in A$. Denote $\bar{p} = p_a$. Axiom 1 implies that

$$\bar{p} u(f_1(1)) + (1-\bar{p}) u(f_0(1)) \le |A|\bar{p} u(f_1(|A|)) + (1 - |A|\bar{p}) u(f_0(|A|)).$$

Using $f_0(k) = 0$ and $f_1(k) = \frac{1}{k}$ for all $k \ge 1$, and that $u(0) = 0$ and $u(1) = 1$ by (3), we obtain

$$\bar{p} \le |A|\bar{p} u\left(\frac{1}{|A|}\right),$$

or $u\left(\frac{1}{|A|}\right) \ge \frac{1}{|A|}$, which is true by (3) and concavity of u.

References

Akeroyd, F. (1982): “Progress in multiple-choice scoring methods,” *Journal of Further and Higher Education*, 6, 86–90.

Alnabhan, M. (2002): “An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric indices of a test,” *Social Behavior and Personality*, 30, 645–652.

Bar-Hillel, M., D. Budescu, and Y. Attali (2005): “Scoring and keying multiple choice tests: A case study in irrationality,” *Mind & Society*, 4, 3–12.

Ben-Simon, A., D. V. Budescu, and B. Nevo (1997): “A Comparative Study of Measures of Partial Knowledge in Multiple-Choice Tests,” *Applied Psychological Measurement*, 21, 65–88.

Bennett, R. E., D. A. Rock, and M. Wang (1991): "Equivalence of Free-Response and Multiple-Choice Items," *Journal of Educational Management*, 28, 77–92.

Bernardo, J. M. (1998): "A Decision Analysis Approach to Multiple-Choice Examinations," in *Applied Decision Analysis*, pp. 195–207. Springer.

Bradbard, D. A., and S. B. Green (1986): "Use of the Coombs elimination procedure in classroom tests," *Journal of Experimental Education*, 54, 68–72.

Bradbard, D. A., D. F. Parker, and G. L. Stone (2004): "An alternative multiple-choice scoring procedure in a macroeconomics course," *Decision Sciences Journal of Innovative Education*, 2, 11–26.

Bridgeman, B. (1991): "Essays and Multiple-Choice Tests as Predictors of College Freshman GPA," *Research in Higher Education*, 32, 319–332.

Brown, R. W. (2001): "Multiple-choice versus descriptive examinations," in *31st ASEE/IEEE Frontiers in Education*. IEEE.

Budescu, D., and M. Bar-Hillel (1993): "To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring," *Journal of Educational Measurement*, 30, 277–291.

Budescu, D. V., and Y. Bo (2014): "Analyzing test-taking behavior: Decision theory meets psychometric theory," *Psychometrika*.

Burton, R. F. (2005): "Multiple-choice and true/false tests: myths and misapprenhensions," *Assessment and Evaluation in Higher Education*, 30(1).

Bush, M. (2001): "A multiple choice test that rewards partial knowledge," *Journal of Further and Higher Education*, 25.

Coombs, C. H., J. E. Miholland, and F. B. Womer (1956): "The assessment of partial knowledge," *Educational and Psychological Measurement*, 16, 13–37.

Delgado, A. R. (2007): "Using the Rasch model to quantify the causal effect of test instructions," *Behavior Research Methods*, 39, 570–573.

Dressel, P. L., and J. Schmidt (1953): "Some modifications of the multiple choice item," *Educational and Psychological Measurement*, 13, 574–595.

Espinosa, M. P., and J. Gardeazabal (2010): "Optimal correction for guessing in multiple-choice tests," *Journal of Mathematical Psychology*, 54, 415–425.

Espinosa, M. P., and J. Gardeazabal (2013): "Do Students Behave Rationally in Multiple Choice Tests? Evidence from a Field Experiment," *Journal of Economics and Management*, 9, 107–135.

Frandsen, G. S., and M. I. Schwartzbach (2006): "A singular choice for multiple choice," in *ACM SIGCSE Bulletin*, vol. 38, pp. 34–38. ACM.

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

Jennings, S., and M. Bush (2006): "A Comparison of Conventional and Liberal (Free-Choice) Multiple-Choice Tests," *Practical Assessment, Research & Evaluation*, 11, 1–5.

Ng, A. W. Y., and A. H. S. Chan (2009): "Different Methods of Multiple-Choice Test: Implications and Design for Further Research," in *Proceedings of IMEC 2009*, vol. II.

Swartz, S. M. (2006): "Acceptance and accuracy of multiple choice, confidence-level, and essay question formats for graduate students," *Journal of Education for Business*, 81, 215–220.

Walstad, W., and W. Becker (1994): "Achievement Differences on Multiple-Choice and Essay Tests in Economics," *American Economic Review*, 84, 193–196.