AARHUS
UNIVERSITY
SCIENCE AND TECHNOLOGY

Singular Choices for Multiple Choice

Department of Computer Science

Olivier Danvy
Martin E. K. Rasmussen

Monograph

Singular Choices for Multiple Choice

Olivier Danvy[†] and Martin E. K. Rasmussen[‡]
Department of Computer Science, Aarhus University[§]

Summer 2016[¶]

Abstract

We revisit and evolve Frandsen and Schwartzbach's axiomatic scoring strategy for multiple-choice exams that credits partial knowledge and levels out guessing. The evolved scoring strategy equalizes the implicit weight of each question by default and makes this weight an optional parameter for each question. Partial credit can also be modulated, which provides a measure of the spread of knowledge of the examinee. Based on a decade of experience in a first-year university course, we find the evolved exams to be more understandable and predictible both for the examiner and for the examinees. Finally, we present a family of scoring functions that fit the model.

[†] Email: `danvy@cs.au.dk`
[‡] Email: `mallereik@gmail.com`
[§] Aabogade 34, DK-8200 Aarhus N, Denmark

[¶] The mention of invisible distractors (Section 3.3.2) was added as this document is going to e-press.

ISBN: 978-87-7507-391-7
DOI: 10.7146/aul.206.149

Contents

1 Background and introduction

Ten years ago, in an article entitled *A Singular Choice for Multiple Choice* [5], Frandsen and Schwartzbach proposed an axiomatic scoring strategy for multiple-choice exams that credits partial knowledge and levels out guessing. They stated six "self-evident" axioms and presented a scoring function that satisfies these axioms. Schwartzbach implemented this scoring function in a service written in Java for generating and scoring multiple-choice tests [14]. An exam is written as an XML document containing grouped questions with multiple answers. The introduction to the exam, the introduction to each group of questions, the text of each question, and the text of each possible answer are embedded as LaTeX fragments. The service can

- generate a LaTeX document containing the multiple-choice exam in the order of the XML document,
- generate arbitrarily many LaTeX documents containing the multiple-choice exam with shuffled groups of questions, shuffled questions, and shuffled possible answers,
- unshuffle and score shuffled actual answers, and
- generate a LaTeX document containing the multiple-choice exam in the order of the XML document where each answer is tagged with the number of students who ticked it (see Figure 1 for a simple example). This document can then be edited further, e.g., with comments (the box in Figure 1), and then published for the benefit of the examinees.

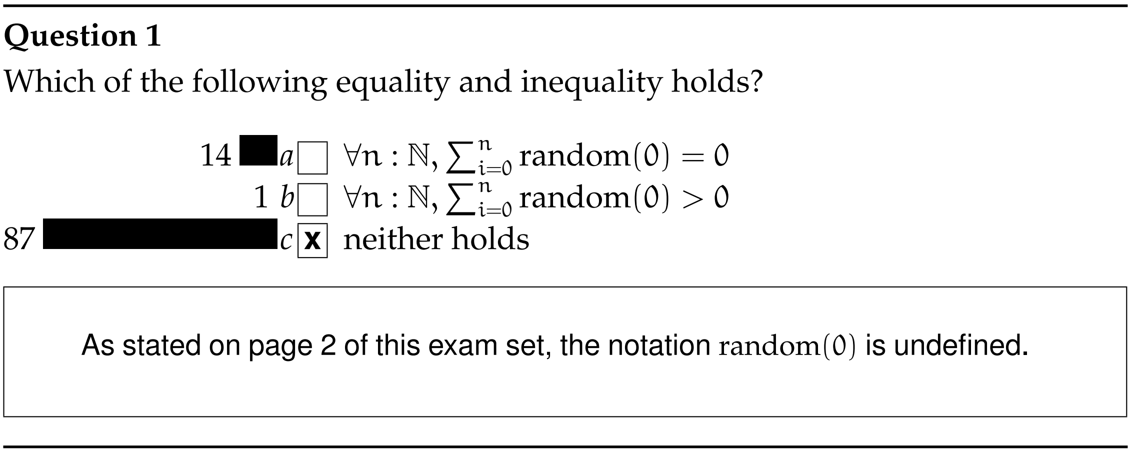
Over the last decade, the first author has used and evolved Schwartzbach's online service for a yearly introductory course targeted to 100+ first-year students. Over the last half-year [12], the second author dedicated his MSc thesis to studying Frandsen and Schwartzbach's singular choice and implementing a simpler, XML-free, and open-source service for generating and scoring multiple-choice tests (see Figure 2 for the counterpart of Figure 1). This service is parameterized by the scoring function to use.

1.1 This article

We show that Frandsen and Schwartzbach's singular choice is not as complex to adopt as it has been presented to be [17] and that it is more than a stepping stone towards other scoring strategies [9, 18]. Our message is that it is actually simple, powerful, and scalable:

- in practice, other styles of multiple-choice exams can be made to fit into this singular choice;
- in theory, the six axioms actually do not determine a unique scoring function, but a family of them, and they themselves are not unique: they can be restated just as sensibly as they were initially stated; the restated axioms determine other families of scoring functions; and
- in principle, axioms about local scoring do not capture all aspects of global scoring, nor do they need to: their true value are that of objective guidelines.

In short, we show that Frandsen and Schwartzbach's contribution was not merely one singular choice for multiple choice, but the first step towards a mathematical framework for specifying singular choices for multiple choice and reasoning about them.



Question 1

Which of the following equality and inequality holds?

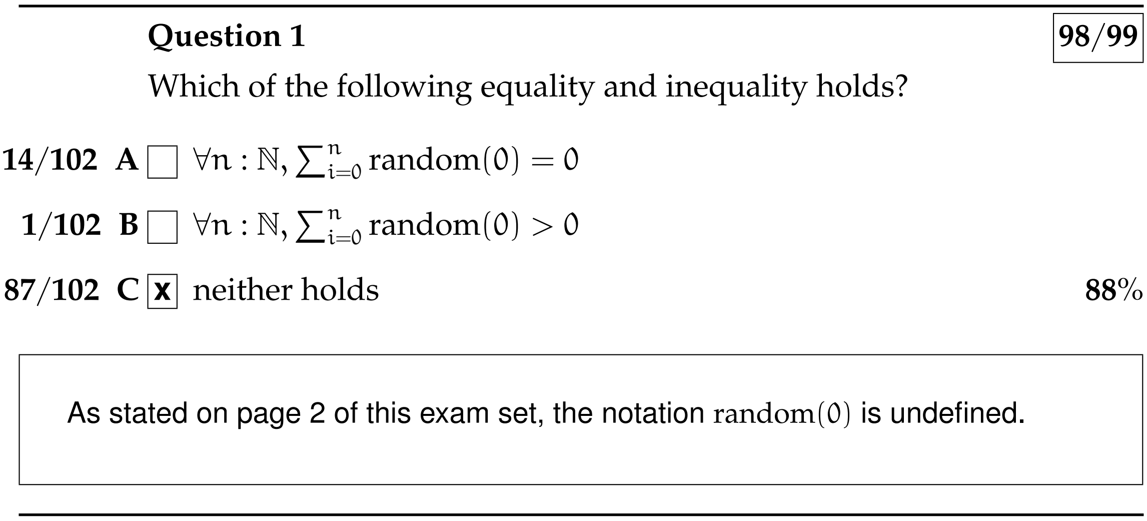
14 *a* □ $\forall n : \mathbb{N}, \sum_{i=0}^{n} \mathrm{random}(0) = 0$

1 *b* □ $\forall n : \mathbb{N}, \sum_{i=0}^{n} \mathrm{random}(0) > 0$

87 *c* ☒ neither holds

As stated on page 2 of this exam set, the notation $\mathrm{random}(0)$ is undefined.

Figure 1: Sample commented question with tagged answers using Schwartzbach's service
The question comes with three possible answers, the two first of which are distractors.
The key was ticked 87 times and the distractors 15 times.
The box contains a subsequent comment from the examiner: for any natural number x, the notation $\mathrm{random}(x + 1)$ stands for a random number between 0 and x.



Question 1 **98/99**

Which of the following equality and inequality holds?

14/102 **A** □ $\forall n : \mathbb{N}, \sum_{i=0}^{n} \mathrm{random}(0) = 0$

1/102 **B** □ $\forall n : \mathbb{N}, \sum_{i=0}^{n} \mathrm{random}(0) > 0$

87/102 **C** ☒ neither holds **88%**

As stated on page 2 of this exam set, the notation $\mathrm{random}(0)$ is undefined.

Figure 2: Sample commented question with tagged answers using the second author's service
Out of 99 examinees, 98 answered this question, 4 of which ticked 2 answers.
This question elicited 102 answers in toto:
A (a distractor) was ticked 14 times, **B** (a distractor) once, and **C** (the key) 87 times.
Overall, $\frac{87}{99} = 88\%$ of the examinees ticked the key.

1.2 Roadmap

We first review Frandsen and Schwartzbach's singular choice for multiple choice, its prior related work, and its subsequent related work (Section 2). Based on a decade of experience using this singular choice for examining 1000+ first-year university students, we then review the issues it raised (Section 3). We propose alternative singular choices for multiple choice (Section 4) and revisit the issues raised in Section 3 (Section 5). We also present a family of scoring functions that fits the model (Section 6) and we revisit Thurstone's scoring function (Section 7).

2 A singular choice for multiple choice

Frandsen and Schwartzbach first specified the axioms a scoring strategy should a priori satisfy, for each question in an exam, in order to credit partial knowledge while leveling out guessing (Section 2.1). Two of these axioms determine the score of a question where the examinee ticked the correct answer (the key) among the other proposed answers (the distractors). Based on this score, another of these axioms determines the score of a question where the examinee did not tick the key. By construction, the resulting local scoring function satisfies the axioms (Section 2.2). A corresponding global scoring function can then be defined that averages the local scoring function, for a given exam set (Section 2.3). We then review related work (Sections 2.4 to 2.6).

2.1 Six "self-evident" axioms

For a question with

- k possible answers (one and only one of which is the key (i.e., the correct answer) and $k - 1$ of which are distractors (i.e., incorrect answers)),
- a actual (ticked) answers, and
- a natural number c indicating the number of correct answers among the actual ones (i.e., 0 or 1),

a scoring strategy S should satisfy the following six "self-evident" axioms. These axioms are parameterized by $\forall k$ such that $k > 1$, $\forall a$ such that $0 \leq a \leq k$:

Axiom 2.1 (No-Answers). $S(k, a, 0) = 0$ *whenever* $a = 0$.
In words: ticking no answers contributes nothing to the score.

Axiom 2.2 (All-Answers). $S(k, a, 1) = 0$ *whenever* $a = k$.
In words: ticking all answers contributes nothing to the score.

Axiom 2.3 (Monotonicity). $S(k, a, 0) < S(k, a, 1)$ *whenever* $0 < a < k$.
In words: ticking the key contributes more to the score than not ticking it.

Axiom 2.4 (Anti-Monotonicity). $\forall a'$ *such that* $0 \leq a' \leq k$, $S(k, a', c) \leq S(k, a, c)$ *whenever* $a \leq a'$.
In words: ticking more distractors incurs more penalty on the score, or in other words: ticking fewer distractors incurs less penalty on the score.

Axiom 2.5 (Invariance). $\forall k'$ *such that* $k' > 1, \forall a'$ *such that* $0 \leq a' \leq k'$, $S(k, a, 1) + S(k', a', 1) = S(k \times k', a \times a', 1)$ *whenever* $0 < a < k$ *and* $0 < a' < k'$.
In words: adding the scores for two questions should yield the same as the score of the conjunction of these two questions and the Cartesian product of their answers.[1]

Axiom 2.6 (Zero-Sum). $\frac{\binom{k-1}{a-1}}{\binom{k}{a}} \cdot S(k, a, 1) + \frac{\binom{k-1}{a}}{\binom{k}{a}} \cdot S(k, a, 0) = 0$ *whenever* $0 < a < k$.
In words: the expected outcome of guessing should be 0*, so that guessing answers contributes nothing to the score, on average.*

In the Zero-Sum axiom, $\binom{k-1}{a-1}/\binom{k}{a}$ is the probability of ticking $a - 1$ distractors and the key, and $\binom{k-1}{a}/\binom{k}{a}$ is the probability of ticking a distractors and not the key. It is simple to verify that $\binom{k-1}{a-1} + \binom{k-1}{a} = \binom{k}{a}$. Furthermore, since $\binom{n}{p} = \frac{n!}{(n-p)! \times p!} = \frac{\prod_{i=1}^{n} i}{\prod_{i=1}^{n-p} i \times \prod_{i=1}^{p} i} = \frac{\prod_{i=p+1}^{n} i}{\prod_{i=1}^{n-p} i} = \frac{(p+1) \times (p+2) \times \ldots \times (n-1) \times n}{1 \times 2 \times \ldots \times (n-p-1) \times (n-p)}$ for $0 \leq p \leq n$, the Zero-Sum axiom simplifies to

$$\frac{a}{k} \cdot S(k, a, 1) + \frac{k-a}{k} \cdot S(k, a, 0) = 0$$

or again:

Axiom 2.7 (Zero-Sum, simplified). $a \cdot S(k, a, 1) + (k - a) \cdot S(k, a, 0) = 0$ *whenever* $0 < a < k$.

2.2 One local scoring function

Here is the local scoring function for a question with k possible answers and a ticked answers:

$$S_{\mathrm{FS}}(k, a, 0) = \begin{cases} 0 & \text{if } a = 0 \\ -\frac{a}{k-a} \cdot \log\left(\frac{k}{a}\right) & \text{if } 0 < a < k \end{cases}$$
$$S_{\mathrm{FS}}(k, a, 1) = \begin{cases} \log\left(\frac{k}{a}\right) & \text{if } 0 < a < k \\ 0 & \text{if } a = k \end{cases}$$

Theorem 2.8 (after the battle). *For any given logarithm base,* S_{FS} *satisfies the six self-evident axioms.*

2.3 One global scoring function

Based on the local scoring function presented in Section 2.2, the global scoring function is obtained by summing the scores for each question and dividing this sum by the sum of the best possible scores:

$$\frac{\sum_i S_{\mathrm{FS}}(k_i, a_i, c_i)}{\sum_i S_{\mathrm{FS}}(k_i, 1, 1)}$$

A negative result indicates that most of the answers were incorrect ones. A non-negative result represents a percentage of positive knowledge, with 1 as 100%. (NB. Frandsen and Schwartzbach max this score with 0 to erase the difference between overall wrong knowledge and no knowledge.)

[1] For example, say Q_1 has 2 possible answers, A_1 and B_1, and Q_2 has 3 possible answers, A_2, B_2, and C_2. The joined question $Q_1 \wedge Q_2$ has $2 \times 3 = 6$ possible answers: $A_1 \wedge A_2$, $A_1 \wedge B_2$, $A_1 \wedge C_2$, $B_1 \wedge A_2$, $B_1 \wedge B_2$, and $B_1 \wedge C_2$. If a answers were ticked in Q_1, including the correct one, and a' answers were ticked in Q_2, including the correct one, then $a \times a'$ are ticked in $Q_1 \wedge Q_2$ and include the correct one.

2.4 Prior related work

Frandsen and Schwartzbach pointed out the variety of related work about multiple-choice exams, from their applicability in higher education (e.g., their coverage) to their psychological dimension (e.g., their understandability and impact). They focused on three references: Bush's work on rewarding partial knowledge [4], Burton's analysis of multiple-choice tests [2], and Roberts's use of multiple-choice tests for formative and summative assessment [13].

2.5 Subsequent related work

Prior to the second author's MSc dissertation [12] and the present article, we found three articles that mention Frandsen and Schwartzbach's work:

- Warwick, Bush, and Jennings [17] mentioned the singular choice but found its format too complex for widespread adoption.
- McCallum [9] described the local scoring function of the singular choice as a whole and then extended it with confidence levels. Indeed, as pointed out in Section 3.2.3, the only indication about confidence in the singular choice is the number of ticks in the answers: the more ticks, the less confidence. However, McCallum did not revisit the original axioms, e.g., to verify whether his extension satisfies them.
- Zapechelnyuk [18] presents a new axiomatization of multiple-choice test scoring, and points out that his axioms are not satisfied by the singular choice.

2.6 Other related work

Multiple-choice testing continues to be an active topic of empirical study [1] and great ingenuity [15], including dissenting voices [6]. For lack of a consensus, alternatives are sought and comprehensively documented [8], with considerable empathy for the examinee [3]. Also, extra demands are made on multiple-choice questions, e.g., formative feedback [11].

There seems to be very little work where basic axioms are sought, and their adequacy, soundness, and completeness are studied.

2.7 Regarding the No-Answers and All-Answers axioms

Originally [5], Axiom No-Answers and Axiom All-Answers were conflated into a Zero axiom:

Axiom 2.9 (Zero). $S(k, k, 1) = S(k, 0, 0) = 0$.
In words: ticking all possible answers is the same as ticking none of them, and contributes nothing to the score.

But the purviews of these two axioms are conceptually distinct: ticking no answers conveys that the examinee does now know which answer may be right, and ticking all answers conveys that the examinee does not know which answer is not wrong. So giving a zero score in both cases is not as self-evident as it was presented to be.

Not ignorance, but ignorance of ignorance, is the death of knowledge.
– Alfred North Whitehead

Here is a rationalization:

The No-Answers axiom It can be seen as the limit of the Anti-Monotonicity axiom for answers to a question where the key is not ticked and no distractors are ticked either. The Anti-Monotonicity axiom says that the fewer distractors are ticked, the less penalty on the score. In the best case, 1 distractor is ticked and the score is $-\frac{1}{k-1} \cdot \log(k)$. The limit of $-\frac{a}{k-a} \cdot \log\left(\frac{k}{a}\right)$ is 0 as a gets close to 0, which is consistent with the No-Answers axiom.

The All-Answers axiom Could we see it as the limit of the Anti-Monotonicity axiom for answers to a question where the key is not ticked and many distractors are ticked? The Anti-Monotonicity axiom says that the more distractors are ticked, the more penalty on the score. In the worst case, all $k-1$ distractors are ticked and the score is $-(k-1) \cdot \log\left(\frac{k}{k-1}\right)$. The limit of $-\frac{a}{k-a} \cdot \log\left(\frac{k}{a}\right)$ is -1 as a gets close to k, which is inconsistent with the All-Answers axiom since if the remaining answer (which happens to be the key) is ticked, the score is 0: a discontinuity.

For the sake of continuity, one could wish to modify the All-Answers axiom to read:

Axiom 2.10 (All-Answers, modified). $S(k, a, 1) = -1$ *whenever* $a = k$.
In words: ticking all possible answers is one notch worse than ticking all the distractors.

With this modified axiom, for a question with k possible answers ($k > 1$) and a ticked answers ($0 \le a \le k$), the local scoring function reads:

$$S'_{\mathrm{FS}}(k, a, 0) = \begin{cases} 0 & \text{if } a = 0 \\ -\frac{a}{k-a} \cdot \log\left(\frac{k}{a}\right) & \text{if } 0 < a < k \end{cases}$$
$$S'_{\mathrm{FS}}(k, a, 1) = \begin{cases} \log\left(\frac{k}{a}\right) & \text{if } 0 < a < k \\ -1 & \text{if } a = k \end{cases}$$

This alternative local scoring function, S'_{FS}, only differs from S_{FS} because of the modified All-Answers axiom.

Theorem 2.11 (after the alternative battle). *For any given logarithm base, S'_{FS} satisfies the six modified self-evident axioms.*

The global scoring function follows mutatis mutandis.

The moral of this subsection is that while S_{FS} was determined by the given self-evident axioms, these axioms are not unique, and modifying one of them leads to another local scoring function, S'_{FS}. In the following section, we explore alternative self-evident axioms and we let them lead us to alternative local scoring functions.

3 Issues about a singular choice

A multiple-choice exam provides a simple and effective bridge from rote learning to critical thinking: it demonstrates that the examinee understands enough of a question and of its possible answers to select its key. At one end of the spectrum, questions and answers make it possible to verify rote learning and to detect confusion (factual and conceptual knowledge [7]). At the other end, questions and answers make it possible not only to invite the examinees to reflect but also to reveal the quality of their reflection (procedural and metacognitive knowledge [7]). A multiple-choice exam enables these assessments independently of how articulate the examinees would be if they were asked to justify their choices or to verbalize their thought process and its outcome, i.e., their reflection.

However, there is what we want to ask the examinees, and what the singular choice offers. This section analyzes the latter and describes the extent to which the former can be made to fit into the latter. For starters (Section 3.1), there is only one correct choice among the options, which openly precludes both "trick questions," i.e., questions for which none of the proposed answers is correct, and questions for which several answers are actually valid. Then there is the issue that both in the model, for the examiner, and for the examinee, all the distractors are equally wrong (Section 3.2), whereas not all the questions weigh as much as each other (Section 3.3). Also, the model assumes all questions to be independent of each other, which forces them to be disconnected (Section 3.4). Finally, the model's locally degressive penalty system makes it somewhat unclear what is globally graded (Section 3.5).

3.1 There is only one correct answer

The singular choice fundamentally assumes that each question has one and only one key. But what about keyless questions (a.k.a. trick questions) and questions that naturally have more than one key?

3.1.1 The glass is half empty

In principle, trick questions go against the grain of a multiple-choice exam: psychologically, the examinees are primed to tick at least one answer, not to refrain from ticking any. In practice, though, trick questions are easily shoehorned into the single-key model by adding an extra possible answer relative to all the others, stating that none of the other options is valid.

3.1.2 The glass is not half full

In principle, having more than one key follows the spirit of a multiple-choice exam, especially one that bestows partial credit for partial knowledge. In practice, though, such questions are not easily shoehorned into the single-key model by adding an extra possible answer relative to all the others, stating, for example, that two of the other options are valid. Such an addition is a red herring since adding this extra possible answer means that there are now three keys, not one. So one is left with duplicating the question so that each duplicate has one key (and distinct distractors), at the risk of confusing the examinee.

3.2 Do the distractors weigh as much as each other? Should they?

3.2.1 In the model

The model assumes the distractors to appear equally likely. The Anti-Monotonicity axiom states that the more distractors are ticked, the less is contributed to the score.

3.2.2 For the examiner

In the mind of the examiner, some distractors are definitely more wrong than others. The model does not account for such degrees of wrongness.

Also, over the years we concluded that multiple-choice exams offer a privileged opportunity to vaccinate examinees against buzzword traps. Before or after vaccination, buzzword traps stand out, which does not fit in the model.

3.2.3 For the examinee

Typically, the examinee has more confidence in some answers than in some others. The model does not account for such degrees of confidence. The only indication it gives is the number of ticks in each question: the more ticks, the less confidence.

3.3 Do the questions weigh as much as each other? Should they?

3.3.1 In the model

By its very definition ($S_{FS}(k, a, 1) = \log\left(\frac{k}{a}\right)$ if $0 < a < k$), the more distractors, the heavier the question in the global score.

3.3.2 For the examiner

In the mind of the examiner, some questions definitely matter more than some others. It would however be a false start to explicitly duplicate a question in the exam proportionally to its weight and warn the examinees about this duplication. Indeed such a duplication introduces noise because the examinees now have the opportunity to hedge their bets and not tick the same answers in each instance of the question, which is a good strategy if only distractors are to be ticked. So, short of

- the ability to declare a weight and to duplicate the question and its answers internally prior to scoring it, or
- the ability to add invisible distractors,

the only viable solution is to comply with the model and add distractors, however weird it feels to convey the importance of a question through the number of ways it can be answered incorrectly.

Let me count the ways.
– Elizabeth Barrett Browning

3.3.3 For the examinee

The examinees need to be explicitly told that questions with the most distractors are the ones that contribute the most to the score.

3.4 Are the questions independent of each other? Should they be?

3.4.1 In the model

The model assumes the questions to be independent of each other. The global scoring function combines the independent scores of each question.

An exam is composed of thematic groups of questions, which mitigates the risk of fragmentation. This risk is real, though: completely independent questions are aligned with partitioned information, and suggest that a disconnected knowledge is enough to pass the exam.

3.4.2 For the examiner

For the examiner, making the questions independent is virtually impossible to achieve because it is antithetical with encouraging the students to reflect. What fits the model, however, is to make *the answers* to each question independent of *the answers* to all the other questions, so that to answer one question, the examinee does not need to first answer another question. And then the examiner is free to continue encouraging the students to reflect, even at the final exam.

CONSTANT VIGILANCE!
– Alastor "Mad Eye" Moody

3.4.3 For the examinee

The examinees need to be explicitly told whether the final exam is aligned with the course and its accretion of knowledge or whether they are better off with a fire-and-forget strategy.

3.5 What is being graded?

Assuming an exam with 100 questions, each with 1 key and 4 distractors, compare

- an examinee who answers 70 questions with 70 ticks; this examinee ticks the key and no distractors for 70 questions, and no key and no distractor for 30 questions;
- an examinee who answers 95 questions with 95 ticks; this examinee ticks the key and no distractors for 75 questions, no key and 1 distractor for 20 questions, and no key and no distractor for 5 questions;
- an examinee who answers 100 questions with 180 ticks; this examinee ticks the key and no distractors for 50 questions, the key and 1 distractor for 20 questions, and the key and 2 distractors for 30 questions; and
- an examinee who answers 100 questions with 145 ticks; this examinee ticks the key and no distractors for 55 questions, the key and 1 distractor for 25 questions, the key and 2 distractors for 10 questions, and no key and 1 distractor for 10 questions.

Tough question: which grade should be given to each of these examinees?

It is tempting to dance around this issue, e.g., using traditional common sense. For an anti-example, the examinee with 70 ticks could either be told that he has a solid base knowledge, is firm in it, and should work on expanding it; or be told that she has a limited knowledge and should work on overcoming her limitations. At the other end, the examinee with 145 ticks could either be told that he displays knowledge and also initiative when facing the unknown, and should work on making this knowledge more firm; or be told that she displays a degree of knowledge but also uncertainty when facing the unknown, and should work on reducing her uncertainties.

Traditional common sense be as it may [10], Frandsen and Schwartzbach's global scoring function gives *the same result* to the four examinees above: 70%. One could wish it to also integrate a dispersion factor among the answers.

4 Singular choices for multiple choice

A forte of the singular choice is that the scoring function is determined by the given axioms. However, as illustrated in Section 2.7, the given axioms are not unique. In that light, to address the tough question of Section 3.5, we introduce a *threshold for partial knowledge*, as a conservative extension of the singular choice (Section 4.1). To make the exam more predictable for the examiner and the examinee, we then make a move toward *equiweighted questions* in any given exam, with and without threshold, again as a conservative extension (Sections 4.2 and 4.3). This move puts us in position to specify an *explicit weight* for each question, again with and without threshold, still as a conservative extension (Sections 4.4 and 4.5).

4.1 A threshold for partial knowledge

Rather than allowing arbitrarily many ticks, one might want to limit their number up to a threshold, as happens often in a web form where not all options can be ticked. For a question with

- k possible answers ($k \in \mathbb{N}$ and $k > 1$) consisting of 1 key and $k - 1$ distractors,
- a threshold t of receivable answers ($t \in \mathbb{N}$ and $0 < t < k$),
- a ticked answers ($a \in \mathbb{N}$ and $0 \leq a \leq k$), and
- a natural number c indicating the number of correct answers among the actual ones ($c \in \mathbb{N}$ and $0 \leq c \leq 1$),

the score of a question is given by $S_{\text{threshold}}(k, t, a, c)$.

The corresponding scoring function should satisfy the following axioms (where the All-Answers axiom has been replaced by a Too-Many-Answers axiom). These axioms are parameterized by $\forall k$ such that $k > 1$, $\forall t$ such that $0 < t < k$, $\forall a$ such that $0 \leq a \leq k$:

Axiom 4.1 (No-Answers). $S_{\text{threshold}}(k, t, a, 0) = 0$ *whenever* $a = 0$ *or* $a > t$.
In words: ticking no answers or too many distractors contributes nothing to the score.

Axiom 4.2 (Too-Many-Answers). $S_{\text{threshold}}(k, t, a, 1) = 0$ *whenever* $a > t$.
In words: above the threshold, ticking answers contributes nothing to the score.

Axiom 4.3 (Monotonicity). $S_{\text{threshold}}(k, t, a, 0) < S_{\text{threshold}}(k, t, a, 1)$ *whenever* $0 < a \leq t$.
In words: within the threshold, ticking the key contributes more to the score than not ticking it.

Axiom 4.4 (Anti-Monotonicity). $\forall a'$ *such that* $0 \leq a' \leq k$, $S_{\text{threshold}}(k, t, a', c) \leq S_{\text{threshold}}(k, t, a, c)$ *whenever* $a \leq a' \leq t$.
In words: within the threshold, ticking fewer distractors incurs less penalty on the score.

Axiom 4.5 (Invariance). $\forall k'$ *such that* $k' > 1$, $\forall t'$ *such that* $0 < t' < k'$, $\forall a'$ *such that* $0 \leq a' \leq k'$, $S_{\text{threshold}}(k, t, a, 1) + S_{\text{threshold}}(k', t', a', 1) = S_{\text{threshold}}(k \times k', t \times t', a \times a', 1)$ *whenever* $0 < a \leq t$ *and* $0 < a' \leq t'$.
In words: within the threshold, adding the scores for two questions should yield the same as the score of the conjunction of these two questions and the Cartesian product of their answers. Logically, the threshold of the compound question should be the product of the thresholds of its components.

Axiom 4.6 (Zero-Sum). $a \cdot S_{\text{threshold}}(k, t, a, 1) + (k - a) \cdot S_{\text{threshold}}(k, t, a, 0) = 0$ *whenever* $0 < a \leq t$.
In words: within the threshold, guessing answers should contribute nothing to the score, on average.

NB. If $t = k-1$, the axioms for $S_{\text{threshold}}$ reduce to the axioms for the original singular choice, S_{FS}.

By the same argument as in Section 2.2, the Invariance axiom and the Anti-Monotonicity axioms determine $S_{\text{threshold}}$ to satisfy $S_{\text{threshold}}(k, t, a, 1) = \log\left(\frac{k}{a}\right)$ for any given logarithm base when a lies within t, and then the Zero-Sum axiom determines $S_{\text{threshold}}$ to satisfy $S_{\text{threshold}}(k, t, a, 0) = -\frac{a}{k-a} \cdot \log\left(\frac{k}{a}\right)$ for the same base when a lies within t. Likewise, the No-answers axiom determines $S_{\text{threshold}}$ to satisfy $S_{\text{threshold}}(k, t, a, 0) = 0$ when a is 0 or is lies outside t and the Too-Many-Answers axiom determines $S_{\text{threshold}}$ to satisfy $S_{\text{threshold}}(k, t, a, 1) = 0$ when a lies outside t.

So all told, here is the local scoring function for a question with k possible answers ($k > 1$), a threshold t ($0 < t < k$), and a ticked answers ($0 \leq a \leq k$):

$$S_{\text{threshold}}(k, t, a, 0) = \begin{cases} 0 & \text{if } a = 0 \\ -\frac{a}{k-a} \cdot \log\left(\frac{k}{a}\right) & \text{if } 0 < a \leq t \\ 0 & \text{if } t < a \end{cases}$$

$$S_{\text{threshold}}(k, t, a, 1) = \begin{cases} \log\left(\frac{k}{a}\right) & \text{if } 0 < a \leq t \\ 0 & \text{if } t < a \end{cases}$$

NB. When $t = k - 1$, there is de facto no threshold, and $S_{\text{threshold}}$ reduces to S_{FS}.

Theorem 4.7 (after the threshold battle). *For any given logarithm base, $S_{\text{threshold}}$ satisfies the six self-evident axioms.*

The global scoring function is obtained by summing the scores for each question and dividing this sum by the sum of the best possible scores:

$$\frac{\sum_i S_{\text{threshold}}(k_i, t_i, a_i, c_i)}{\sum_i S_{\text{threshold}}(k_i, t_i, 1, 1)}$$

A negative result indicates that most of the answers were incorrect ones. A non-negative result represents a percentage of positive knowledge, with 1 as 100%.

Again, there is room for variation here:

- For example, the threshold could be a fraction of the number of actual answers over the number of possible answers, e.g., half or a third.

- For example, the No-Answers axiom and the Too-Many-Answers axiom could be tuned so that $S_{\text{threshold}}(k, t, a, c)$ is one notch worse when a exceeds t—though which examinee would tick answers above the threshold? Still, one would then need to tinker with the Zero-Sum axiom to level out guessing when the number of ticked answers exceeds the threshold, which would lead to another local scoring function.

4.2 Equiweighting questions

Section 3.3 pointed out that the weight of each question is proportional to its number of possible answers, k. Let k_{max} be the largest number of possible answers in all the questions of a given exam. It is simple to give the same weight to each of the questions in this given exam: map each question with k possible answers to a question with k_{max} possible answers and a threshold of $k-1$. This way, all the questions will be treated as if each had k_{max} possible answers (and thus they all are equiweighted) and a threshold of receivable answers that corresponds to its actual number of possible answers, as per the note just before Theorem 4.7.

So for a question with k possible answers ($k > 1$) and a ticked answers ($0 \leq a \leq k$), the local scoring function in an exam where the largest number of possible answers is k_{max} reads as follows:

$$S^{\text{equalized}}(k, a, 0) = S_{\text{threshold}}(k_{max}, k-1, a, 0) = \begin{cases} 0 & \text{if } a = 0 \\ -\frac{a}{k_{max}-a} \cdot \log\left(\frac{k_{max}}{a}\right) & \text{if } 0 < a < k \end{cases}$$

$$S^{\text{equalized}}(k, a, 1) = S_{\text{threshold}}(k_{max}, k-1, a, 1) = \begin{cases} \log\left(\frac{k_{max}}{a}\right) & \text{if } 0 < a < k \\ 0 & \text{if } a = k \end{cases}$$

All the questions now have the same maximal weight in the global score, namely $\log(k_{max})$. As such, $S^{\text{equalized}}$ does not satisfy six self-evident axioms: it is defined in terms of a scoring function that does, $S_{\text{threshold}}$.

NB. $S^{\text{equalized}}$ reduces to S_{FS} whenever all the questions have the same number of possible answers.

For this given exam, we can compute a global score as usual, i.e., by summing the scores for each question and dividing this sum by the sum of the best possible scores:

$$\frac{\sum_i S^{\text{equalized}}(k_i, a_i, c_i)}{\sum_i S^{\text{equalized}}(k_i, 1, 1)}$$

or again, if the exam contains n questions,

$$\frac{\sum_{i=1}^{n} S_{\text{threshold}}(k_{max}, k_i - 1, a_i, c_i)}{n \cdot \log(k_{max})}$$

A negative result indicates that most of the answers were incorrect ones. A non-negative result represents a percentage of positive knowledge, with 1 as 100%.

With this equalizing global scoring function, an examinee who ticks j% of the keys and 0 distractors gets a score of j%, which concurs with traditional intuition both for the examiner (who needs to compose the exam as sensibly as possible) and for the examinee (who needs to pass this exam as unobtrusively as possible). Besides, the questions in the exam can now have as many options as the examiner sees fit, without affecting their weight.

4.3 A threshold for partial knowledge in equiweighted questions

As a simple corollary of Sections 4.1 and 4.2, we can have both a threshold of receivable answers and equiweighted questions by specifying our own threshold, t, rather than the largest possible one, $k - 1$.

For a question with k possible answers ($k > 1$), a threshold t ($0 < t < k$), and a ticked answers ($0 \leq a \leq k$), the local scoring function in an exam where the largest number of possible answers is k_{max} reads as follows:

$$S^{\text{equalized}}_{\text{threshold}}(k, t, a, 0) = S_{\text{threshold}}(k_{max}, t, a, 0) = \begin{cases} 0 & \text{if } a = 0 \\ -\frac{a}{k_{max}-a} \cdot \log\left(\frac{k_{max}}{a}\right) & \text{if } 0 < a \leq t \\ 0 & \text{if } t < a \end{cases}$$

$$S^{\text{equalized}}_{\text{threshold}}(k, t, a, 1) = S_{\text{threshold}}(k_{max}, t, a, 1) = \begin{cases} \log\left(\frac{k_{max}}{a}\right) & \text{if } 0 < a \leq t \\ 0 & \text{if } t < a \end{cases}$$

Again, $S^{\text{equalized}}_{\text{threshold}}$ does not satisfy the self-evident axioms: it is defined in terms of a scoring function that does, $S_{\text{threshold}}$.

NB. $S^{\text{equalized}}_{\text{threshold}}$ reduces to $S^{\text{equalized}}$ whenever $t = k-1$, and to S_{FS} whenever all the questions have the same number of possible answers, k_{max}, and the same threshold, $k_{max} - 1$.

For this given exam, we can compute a global score as usual, i.e., by summing the scores for each question and dividing this sum by the sum of the best possible scores:

$$\frac{\sum_i S^{\text{equalized}}_{\text{threshold}}(k_i, t_i, a_i, c_i)}{\sum_i S^{\text{equalized}}_{\text{threshold}}(k_i, t_i, 1, 1)}$$

or again, if the exam contains n questions,

$$\frac{\sum_{i=1}^{n} S_{\text{threshold}}(k_{max}, t_i, a_i, c_i)}{n \cdot \log(k_{max})}$$

A negative result indicates that most of the answers were incorrect ones. A non-negative result represents a percentage of positive knowledge, with 1 as 100%.

4.4 Specifying a weight for each question

Now that each question implicitly weighs the same in the global score, we can address the point made in Section 3.3, namely that some questions should have more weight than others, and should be explicitly declared as such. So let us add another facet to a question: its weight, $w : \mathbb{N}$. In Section 3.3.2, it was the wish of the examiner to have the ability to duplicate the question and its answers in proportion of its weight, prior to scoring it: 0 to not consider it, 1 to not duplicate it, 2 to make two copies, etc. It therefore seems pretty self-evident, so to speak, that given a question with weight w, k possible answers, a actual answers, and a natural number c indicating whether the key was ticked, the local scoring strategy S^{weight} should be defined to satisfy

$$\begin{aligned} S^{\text{weight}}(w, k, a, c) &= \underbrace{S^{\text{equalized}}(k, a, c) + \cdots + S^{\text{equalized}}(k, a, c)}_{w \text{ times}} \\ &= w \times S^{\text{equalized}}(k, a, c) \end{aligned}$$

So for a question of weight w ($w > 0$) with k possible answers and a ticked answers, the local scoring function reads as follows:

$$\begin{aligned} S^{\text{weight}}(w, k, a, 0) = w \times S^{\text{equalized}}(k, a, 0) &= \begin{cases} 0 & \text{if } a = 0 \\ -\frac{a}{k-a} \cdot w \cdot \log\left(\frac{k}{a}\right) & \text{if } 0 < a < k \end{cases} \\ S^{\text{weight}}(w, k, a, 1) = w \times S^{\text{equalized}}(k, a, 1) &= \begin{cases} w \cdot \log\left(\frac{k}{a}\right) & \text{if } 0 < a < k \\ 0 & \text{if } a = k \end{cases} \end{aligned}$$

Again, S^{weight} does not satisfy the self-evident axioms: it is defined in terms of a scoring function that does, $S_{\text{threshold}}$.

NB. S^{weight} reduces to $S^{\text{equalized}}$ whenever the question has weight one ($w = 1$). A question with weight zero ($w = 0$) contributes nothing to the score, which is practical a posteriori if it contained a typo or if it was massively misunderstood by the examinees.

The global scoring function is obtained by summing the scores for each question and dividing this sum by the sum of the best possible scores:

$$\frac{\sum_i S^{\text{weight}}(w_i, k_i, a_i, c_i)}{\sum_i S^{\text{weight}}(w_i, k_i, 1, 1)}$$

A negative result indicates that most of the answers were incorrect ones. A non-negative result represents a percentage of positive knowledge, with 1 as 100%.

4.5 Specifying a weight for each question with a threshold for partial knowledge

As a simple corollary of Sections 4.1 and 4.4, we can specify both the weight of each individual question and a threshold of receivable answers for each question.

For a question with k possible answers ($\mathrm{k} > 1$), a threshold t ($0 < \mathrm{t} < \mathrm{k}$), and a ticked answers ($0 \leq \mathrm{a} \leq \mathrm{k}$), the local scoring function in an exam where the largest number of possible answers is k_{max} reads as follows:

$$S^{\text{weight}}_{\text{threshold}}(w, k, t, a, 0) = w \times S^{\text{equalized}}_{\text{threshold}}(k, t, a, 0) = \begin{cases} 0 & \text{if } a = 0 \\ -\frac{a}{k_{max}-a} \cdot w \cdot \log\left(\frac{k_{max}}{a}\right) & \text{if } 0 < a \leq t \\ 0 & \text{if } t < a \end{cases}$$

$$S^{\text{weight}}_{\text{threshold}}(w, k, t, a, 1) = w \times S^{\text{equalized}}_{\text{threshold}}(k, t, a, 1) = \begin{cases} w \cdot \log\left(\frac{k_{max}}{a}\right) & \text{if } 0 < a \leq t \\ 0 & \text{if } t < a \end{cases}$$

Again, $S^{\text{weight}}_{\text{threshold}}$ does not satisfy the self-evident axioms: it is defined in terms of a scoring function that does, $S_{\text{threshold}}$.

NB. $S^{\text{weight}}_{\text{threshold}}$ reduces to $S^{\text{equalized}}_{\text{threshold}}$ whenever $w = 1$.

For this given exam, we can compute a global score as usual, i.e., by summing the scores for each question and dividing this sum by the sum of the best possible scores:

$$\frac{\sum_i S^{\text{weight}}_{\text{threshold}}(w_i, k_i, t_i, a_i, c_i)}{\sum_i S^{\text{weight}}_{\text{threshold}}(w_i, k_i, t_i, 1, 1)}$$

A negative result indicates that most of the answers were incorrect ones. A non-negative result represents a percentage of positive knowledge, with 1 as 100%.

5 Issues about singular choices

This section revisits Section 3 in the light of Section 4, comparing and contrasting what we want to ask the examinees and what the singular choices offer. We first scrutinize the unicity of the key among the options (Section 5.1). We then revisit the issues of the relative wrongness of all the distractors (Section 5.2), the relative weight of all the questions (Section 5.3), the independence of all the questions (Section 5.4), and what is globally graded (Section 5.5).

5.1 There is only one correct answer

The singular choices fundamentally assume that each question has one and only one key. But what about keyless questions and questions that naturally have more than one key?

5.1.1 The glass is still half empty

"Trick questions" (i.e., questions where none of the suggested answers is correct and therefore where the right answer is to not tick any of the suggested answers) would require a new model. We still believe that they go against the grain of a multiple-choice exam, where psychologically, the examinees are primed to tick answers, not to leave answers blank. Since trick questions are easily shoehorned with an "else" option, we conclude that they raise no new issues here.

5.1.2 The glass is still not half full

Having more than one key would still require a new model.

5.2 Do the distractors weigh as much as each other? Should they?

As stated, the singular choices do not affect the relative weights of the distractors, and the same issues as in Section 3.2 apply. One would need to expand the model with dual degrees of wrongness (for the examiner) and of confidence (for the examinee). This expansion is not straightforward if the Anti-Monotonicity axiom is to still hold.

Also, in this context, wrongness is mature and absolute whereas confidence is relative and evolving: the examiner has an total grasp of the material for a long time whereas examinees are newcomers who are acquiring a new knowledge and who are building confidence as they go. In that sense, asking first-year students about their confidence seems premature.

5.3 Do the questions weigh as much as each other? Should they?

As developed in Section 4.2, with the advent of $\mathtt{k}_{max}$ in the global scoring function, questions can now have the same weigh as each other. Furthermore, as developed in Section 4.4, the examiner can now declare a weight for each question (0 means that the question does not count in the final score).

Of course, questions with a non-standard weight should be singled out, for the examinee's information.

5.4 Are the questions independent of each other? Should they be?

The singular choices bring nothing new to the matter of Section 3.4.

5.5 What is being graded?

The combination of equiweighted questions and the option for the examiner to give individual weights for each question provides for a more balanced estimate of the global score. Based on a number of random exam copies with equiweighted questions and on simulating various distributions of answers, we have verified that (1) guessing answers contributes nothing to the score, on average, (2) total (resp. partial) knowledge gives a total (resp. partial) score, and (3) in case of contra-positive knowledge, it pays to tick multiple answers rather than none at all:

(1a) An exam with random answers gets a 0% score on average.

(1b) An exam where 3/4 of the questions were answered with a single tick and the key was ticked 2 times out of 3 gets a 44% score. Guessing single ticks in the remaining quarter elicits an average score of 44%. Double ticks, 44%.

(1c) An exam where 2/3 of the questions were answered with a single tick and the key was ticked 3 times out of 4 gets a 46% score. Guessing single ticks in the remaining third elicits an average score of 46%. Double ticks, 46%.

(2) An exam where 9/10 of the questions were correctly answered with a single tick gets a 90% score. With 2 ticks, a 51% score. With 3 ticks, a 29% score.

(3) An exam where 1/2 of the questions were answered correctly with a single tick gets a 50% score. If 2 distractors out of 5 can be eliminated in the remaining questions, guessing single ticks elicits an average score of 58%. Double ticks, 63%. Ticking all 3 options, 66%. So if the examinee knows enough to correctly eliminate distractors in a question but not enough to correctly select the key among the remaining possible answers, it pays to tick them all.

Playing with the threshold makes it possible to identify the dispersion factor among the answers, i.e., the standard deviation. For example, in Section 3.5, the global score of the examinee with 180 ticks goes from 70% to 61% with a threshold of 2. Also, a threshold of 1 indicates how many questions were correctly answered with one tick. In other words, the variety of scoring functions makes it possible to data mine each copy.

Of course, if a question has a different weight compared to the others, and if the threshold affects the score, the examinee should be told so.

6 More singular choices

The goal of this section is to derive a family of local scoring functions S that generalize S_{FS} and that all satisfy Frandsen and Schwartzbach's self-evident axioms. We start from the axioms in Section 2.1, i.e., Axioms 2.1 to 2.6, keeping in mind that $k > 1$, $0 \leq a < k$ if $c = 0$, and $0 < a \leq k$ if $c = 1$. The method is as follows:

- we start from a definition $S(k, a, 1) = X$ that satisfies the Invariance axiom;
- we use the Zero-Sum axiom to define $S(k, a, 0) = -\frac{a}{k-a} \cdot X$;
- we verify whether the Monotonicity and Anti-Monotonicity axioms are satisfied; and
- if that is the case, we use the No-Answers axiom and the All-Answers axiom to define the following local scoring function for a question with k possible answers and a ticked answers:

$$S(k, a, 0) = \begin{cases} 0 & \text{if } a = 0 \\ -\frac{a}{k-a} \cdot X & \text{if } 0 < a < k \end{cases}$$
$$S(k, a, 1) = \begin{cases} X & \text{if } 0 < a < k \\ 0 & \text{if } a = k \end{cases}$$

By construction, this function satisfies the six self-evident axioms.

The Invariance axiom requires that $S(k, a, 1) = \log(f(k, a))$, for some function f and for any given logarithm base, and this function must satisfy

$$\begin{cases} f(k, a) & \neq & 1 \\ \log(f(k, a)) + \log(f(k', a')) & = & \log(f(k \times k', a \times a')) \end{cases}$$

or again, since $\log(f(k, a)) + \log(f(k', a')) = \log(f(k, a)) \times (f(k', a'))$,

$$\begin{cases} f(k, a) & \neq & 1 \\ f(k, a) & > & 0 \\ f(k, a) \times f(k', a') & = & f(k \times k', a \times a'). \end{cases}$$

So the only possible candidates for $f(k, a)$ are products of powers of k and of a, i.e.,

$$S(k, a, 1) = \log(k^p \times a^{p'})$$

or again

$$S(k, a, 1) = p \cdot \log k + p' \cdot \log a$$

for any real numbers p and p'. By the Zero-Sum axiom, $S(k, a, 0) = -\frac{a}{k-a} \cdot \log\,(k^p \times a^{p'})$.

The condition $p' \leq 0$ must hold for the Anti-Monotonicity axiom to be satisfied, and the conditions $p > 0$ and $p \geq -p'$ for the Monotonicity axiom to be satisfied.

So all told, for any real coefficients $p > 0$ and $q \geq 0$ such that $p \geq q$, the following local scoring function satisfies the axioms, by construction:

$$S(k, a, 0) = \begin{cases} 0 & \text{if } a = 0 \\ -\frac{a}{k-a} \cdot \log\,(\frac{k^p}{a^q}) & \text{if } 0 < a < k \end{cases}$$
$$S(k, a, 1) = \begin{cases} \log\,(\frac{k^p}{a^q}) & \text{if } 0 < a < k \\ 0 & \text{if } a = k \end{cases}$$

Letting both p and q be 1 gives S_{FS}, and letting q be 0 does not penalize partial knowledge if the key was ticked (it does, however, penalize partial ignorance). Extending this family of functions with a threshold and with equiweighted as well as individually weighted questions is done along the lines of Section 4. For the rest, we do not have a sense of what it means to play with p and q in practice—a future work.

7 Why a new scoring function?

The grandmother of all scoring functions is nearly a century old, and is due to Thurstone [16]. In closed form, it reads

$$S_T(k, a, c) = c - \frac{a - c}{k - 1}$$

for a multiple-choice question with k possible answers, a actual answers, and c keys.

Except for the Invariance axiom, Thurstone's scoring function satisfies all of Frandsen and Schwartzbach's self-evident axioms when $c = 1$:

No-Answers $S_T(k, 0, 0) \overset{\text{def}}{=} 0 - \frac{0-0}{k-1} = 0$, and so ticking no answers contributes nothing to the score;

All-Answers $S_T(k, k, 1) \overset{\text{def}}{=} 1 - \frac{k-1}{k-1} = 0$, and so ticking all answers contributes nothing to the score;

Monotonicity $S_T(k, a, 0) \overset{\text{def}}{=} 0 - \frac{a-0}{k-1} = -\frac{a}{k-1} < \frac{k}{k-1} - \frac{a}{k-1} = 1 + \frac{1}{k-1} - \frac{a}{k-1} = 1 - \frac{a}{k-1} + \frac{1}{k-1} = 1 - \frac{a-1}{k-1} \overset{\text{def}}{=} S_T(k, a, 1)$, and so ticking the key contributes more to the score than not ticking it;

Anti-Monotonicity $\forall a'$ such that $0 \leq a' \leq k$ and $a \leq a'$ (i.e., $-a' \leq -a$), $S_T(k, a', c) \overset{\text{def}}{=} c - \frac{a'-c}{k-1} = c - \frac{a'}{k-1} + \frac{c}{k-1} \leq c - \frac{a}{k-1} + \frac{c}{k-1} = c - \frac{a-c}{k-1} \overset{\text{def}}{=} S(k, a, c)$, and so ticking fewer distractors incurs less penalty on the score; and

Zero-Sum $a \cdot S_T(k, a, 1) + (k - a) \cdot S_T(k, a, 0) \overset{\text{def}}{=} a \cdot (1 - \frac{a-1}{k-1}) + (k - a) \cdot (0 - \frac{a-0}{k-1}) = \frac{a \cdot k - a - a^2 + a + k \cdot a - a^2}{k-1} = 0$ and so the expected outcome of guessing is 0.

So the key point of the singular choice lies in the Invariance axiom. This axiom captures that all the questions, in an exam, are independent of each other, a point which is orthogonal to Thurstone's scoring function.

As analyzed in Section 3.4, however, this independence is not above criticism: like the proverbial unappetizing soup whose portions are deemed too small, independence is difficult to achieve and also it is not desirable in general. What is needed here is (1) a dependence graph for each exam, (2) the corresponding graph of strongly connected components, and (3) instances of the Independence axiom that are relative to this other graph – a future work.

Another issue is that Thurstone's scoring function naturally accounts for multiple keys, whereas for information-theoretic reasons, Frandsen and Schwartzbach consider that handling multiple keys in their singular choice is an open problem.

On the other hand, Thurstone's scoring function lends itself readily to a global scoring function that sums the scores for each question and divides this sum by the sum of the best possible scores:

$$\frac{\sum_i S_T(k_i, a_i, c_i)}{\sum_i S_T(k_i, c_i, c_i)}$$

With such a global scoring function,

- questions are equiweighted by default, independently of their number of options and keys (Section 4.2);
- questions are easy to weigh individually (Section 4.4); and
- it is simple to specify a threshold for partial knowledge (Section 4.3).

And here is where the axiomatic approach pays off: we are not doomed to only subjectively post-justify each of these properties, variations, and extensions, e.g., with empirical field tests – we can also verify whether the axioms hold, objectively.

8 Conclusion and perspectives

For the examiners, a chief advantage of a multiple-choice exam is that most of their overall time allocated to an exam is dedicated to composing this exam, not to correcting it and having to justify each correction. For the first-year examinees, multiple-choice exams educates them to understand questions in order to answer them. However, designing satisfactory scoring strategies seems as complicated as designing satisfactory voting strategies: they need to render an accurate picture of the knowledge of the examinee, they need to be conducive to the rendering of this picture, they need to be aligned with the course material, they need to be invulnerable to the circumstances of the exam, and they need to lead to exams that are as simple as possible both for the examiner and for the examinees so that both can focus on the content of the exam rather than struggle with its form.

Ten years ago, Frandsen and Schwartzbach turned to mathematics to design a scoring strategy, stated "self-evident" axioms, and derived a scoring function. Such an axiomatic approach is characterized as follows:

Adequacy Do the constitutive axioms adequately account for what is wanted of a scoring strategy?

Soundness How independent are the axioms from each other? Do they, in combination, only lead to properties or characterizations that are compatible or even supportive of what is wanted in a scoring strategy?

Completeness Is everything that is wanted of a scoring strategy accounted for by the axioms?

Frandsen and Schwartzbach's take on adequacy was that their axioms are self-evident, their take on soundness is that the resulting scoring function compares to pre-existing ones, and they do not mention completeness. The forte of their singular choice is that the resulting scoring function is derived, not guesstimated or empirically analyzed, and that it is internally coherent, provided that the axioms do not contradict each other.

In this article, we have revisited and evolved Frandsen and Schwartzbach's singular choice in the light of a decade of experience using it for the final exam of a first-year introductory course about programming languages:

Adequacy We have introduced the notions of threshold for partial knowledge, of controllable weight for individual questions, and of standard deviation to acquire a sense of the firmness of the examinee's knowledge.

Soundness We have refined the axioms to make them independent of each other. (The original formulation had a minor overlap.) We have shown that the axioms are not a straitjacket: they can support a variety of concerns (e.g., trick questions). Also, they are not unique.

Completeness Such an axiomatization is limited to the technical aspect of scoring strategies. For example, it does not account for making the previous exams available as a preparation factor for the examinees, for shuffling questions and possible answers in exam copies, etc.

We have also presented a family of scoring functions that fit the model.

Overall, designing satisfactory scoring strategies involves such a variety of concerns that it makes sense to look for objective, verifiable guidelines that can be reasoned with, hence the present axiomatic approach.

Acknowledgments Gudmund Frandsen patiently provided multiple answers to a series of questions about the singular choice. Irene Danvy and Jacob Johannsen offered timely comments as well as perspective. Karoline Malmkjær raised the tough question of Section 3.5. Mayer Goldberg was a fountain of wisdom, zeroed on the formula $p \cdot \log k + p' \cdot \log a$ as the most general solution for the Invariance axiom, in Section 6, and pointed out that the coefficients were real numbers, not just integers.

References

[1] Lucy R. Betts, Tracey J. Elder, James Hartley, and Mark Trueman. Does correction for guessing reduce students' performance on multiple-choice examinations? yes? no? sometimes? *Assessment & Evaluation in Higher Education*, 34(1):1–15, 2009.

[2] Richard F. Burton. Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education*, 30(1):65–72, February 2005.

[3] Richard F. Burton. Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education*, 40(2):218–231, 2015.

[4] Martin Bush. A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education*, 25(2):157–163, 2001.

[5] Gudmund S. Frandsen and Michael I. Schwartzbach. A singular choice for multiple choice. *ACM SIGCSE Bulletin*, 32(4):34–38, December 2006.

[6] Richard B. Gunderman and Joseph M. Ladowski. Inherent limitations of multiple-choice testing. *Academic Radiology*, 20(10):1319–1321, October 2013.

[7] David R. Krathwohl. A revision of Bloom's taxonomy: an overview. *Theory into Practice*, 41(4):212–218, 2002.

[8] Ellen Lesage, Martin Valcke, and Elien Sabbe. Scoring methods for multiple choice assessment in higher education – is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation*, 39:188–193, 2013.

[9] Simon McCallum. Game design for computer science education. In Erik Hjelmås, editor, *Proceedings of the 2010 Norsk Informatikkonferanse (NIK 2010)*, Gjøvik, Norway, November 2010. Session 3: Pedagogikk og utdanning.

[10] Lene Mejlby. Why are there so few female students in computer science? Master's thesis, Department of Computer Science, Aarhus University, Aarhus, Denmark, June 2010.

[11] Andrew Petersen, Michelle Craig, and Paul Denny. Employing multiple-answer multiple choice questions. In Janet Carter and Yvan Tupac, editors, *ITiCSE '16: Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*, pages 252–253, Arequipa, Peru, July 2016. ACM.

[12] Martin E. K. Rasmussen. A flat approach to syntax-checking, shuffling, and correcting multiple-choice tests. Master's thesis, Department of Computer Science, Aarhus University, Aarhus, Denmark, June 2016.

[13] Tim S. Roberts. The use of multiple choice tests for formative and summative assessment. In Denise Tolhurst and Samuel Mann, editors, *Eighth Australasian Computing Education Conference (ACE 2006)*, volume 52 of *Conferences in Research in Practice in Information Technology*, pages 175–180, Hobart, Tasmania, Australia, January 2006. Australian Computer Society.

[14] Michael Schwartzbach. Multiple choice tool. `http://users-cs.au.dk/mis/Multiple/`, November 2006.

[15] Simon. Wrong is a relative concept: part marks for multiple-choice questions. In John Hamer and Michael de Raadt, editors, *Thirteenth Australasian Computing Education Conference (ACE 2011)*, volume 114 of *Conferences in Research in Practice in Information Technology*, pages 47–53, Perth, Australia, 2011. Australian Computer Society.

[16] Louis Leon Thurstone. A scoring method for mental tests. *Psychological Bulletin*, 17:235–240, 1919.

[17] Jon Warwick, Martin Bush, and Sylvia Jennings. Analysis and evaluation of liberal (free-choice) multiple-choice tests. *Innovation in Teaching and Learning in Information and Computer Sciences*, 9(2):1–12, 2010.

[18] Andriy Zapechelnyuk. An axiomatization of multiple-choice test scoring. *Economics Letters*, 132:24–27, 2015.