Item Order and Speededness: Implications for Test Fairness in Higher Educational High-Stakes Testing

Abstract

A common approach to increase test security in higher educational high-stakes testing is the use of different test forms with identical items but different item orders. The effects of such varied item orders are relatively well studied, but findings have generally been mixed. When multiple test forms with different item orders are used, we argue that the moderating role of speededness on item order effects cannot be neglected as missing responses are commonly scored as incorrect in high-stakes testing. If test-takers run out of time while not giving answers to easy items at the end of the test, they are penalized stronger than if instead they were unable to provide answers to difficult items. Using an illustrative real-data example of a speeded test, we show that the potential consequences of ignoring item order can be substantial with respect to test fairness. Our proposed solution consists of using a fixed item order across forms from the point at which the test may become speeded for some students. In this approach, the most time-intensive items are placed at the end of the test. A simulation based on real data of two university exams from psychology students illustrates the usefulness of this approach.

In higher educational high-stakes tests like college exams, important challenges are test fairness and test security. To assure test fairness, it is popular practice to set a common time limit for all test-takers and to score missing responses as incorrect, to prevent test-takers from choosing a specific set of items to respond to. With regard to test security, a major concern is cheating and more specifically, test-takers copying answers from other test-takers, the most popular cheating practice in crowded class room situations (Chirumamilla et al., 2020). A common approach to prevent this behavior is to create multiple test forms with rearranged item orders and to provide neighboring test-takers with differentially ordered forms (e.g., Monk & Stallings, 1970). This strategy is assumed to limit the probability that neighboring test-takers are simultaneously working on the same items, thereby making answer-copying difficult (Davis, 2017; Vander Schee, 2013). Other methods that prevent answer-copying, like test forms with distinct item sets or computer adaptive testing exist (van der Linden, 2005), typically require larger item pools, pretesting items, and/or computer-based test administrations. However, these requirements are often impossible to meet in conventional higher educational testing.

If multiple test forms with different item orderings are used, the resulting test scores should not depend on the ordering of the items. Or, as Lord (1980, p.195) writes, "..., it must be a matter of indifference to applicants at every given ability level [...] whether they are to take test x or test y". A test cannot be considered a fair test, if the test score of an individual would be different given an alternative test form.

In this paper, we investigate how different item orderings can affect test performance for test-takers and therefore violate principles of test fairness. First, we give a brief overview of the research on item order effects. Then, we introduce a modeling framework which allows us to jointly model ability and speed. Based on this framework, we introduce the concept of *speededness* and discuss why speededness may have been (partly) overlooked when explaining item order effects. Furthermore, test-takers can act in different manners when facing speededness constraints on a test. Using the concept of test-wiseness

we discuss these differences and explain how test-wiseness influences the relationship between speededness and test fairness. An empirical example of a speeded test is used to demonstrate how different item orderings can lead to unfair test forms. Finally, we propose a simple, heuristic approach to prevent unfair effects of item ordering and illustrate its effectiveness based on a short simulation study.

Theoretical Background

The question whether item order can be rearranged without affecting the fairness of a test has been extensively discussed in the literature. Leary and Dorans (1985) provide an exhaustive overview of the research before 1985, whereas Wang (2019) provides a more recent, but smaller overview. Most studies have focused on whether overall test difficulty varies if items are sorted (a) in random order, (b) Easy-Hard, (c) Hard-Easy, or (d) ordered according to content (*topical ordering*). Note that all specific orderings (such as b-d) can also result from random ordering. Therefore, even if test administrators plan to use random item orderings, they have to ensure that any possible differential impact on test scores due to item order is avoided.

Leary and Dorans (1985) state that sorting by difficulty usually has an effect on test scores if the test is administered under a time limit, with Hard-Easy leading to the lowest scores. They offer the explanation that in the Hard-Easy conditions "...when an examination is administered under strict time constraints, some examinees could be at a disadvantage as a result of spending time on hard items early in the test that they could more profitably have spent on easy items near the end". For a similar explanation see also Sax and Cromack (1966), who conclude that "... test constructors have a responsibility of arranging items in ascending order of difficulty if tests are lengthy or time limits restricted"[1].

[1] Note that this would not be the case if missing responses were not scored as incorrect. For example in large-scale low-stakes assessments there is a vivid discussion revolving around alternative scoring or modeling techniques (e.g., Rose et al., 2017). However, due to reasons of test fairness these approaches are hardly applicable to high-stakes testing.

Overall, however, the literature is inconclusive regarding the relation between item orderings and test difficulty: Leary and Dorans (1985) report contradicting findings; a meta analysis by Aamodt and McShane (1992) reports small but significant effects; some more recent studies find no effects (Chidomere, 1989; Davis, 2017; Neely et al., 1994; Perlini et al., 1998; Vander Schee, 2013) while other recent studies do find difficulty differences across different item orderings (Chen, 2012; Pettit et al., 1986; Russell et al., 2003; Togo, 2002).

Although not aimed to explain these mixed results, studies have explored different aspects that can play a role in the relation between item order and test performance. For instance, the role of test anxiety, either as a moderator (if test anxiety is viewed as a trait, Chen, 2012) or a mediator (if test anxiety is viewed as a state, McKeachie et al., 1955) has been investigated. Further, the impact of topically ordered items on ease of memory retrieval has been studied (Pettit et al., 1986; Togo, 2002). In this paper, however, we focus on the role of test speededness. More specifically, we address the hypothesis stated by Leary and Dorans (1985) and Sax and Cromack (1966) above: If a test is administered under time constraints, different item orderings can substantially and differentially affect the test scores of individuals as it leads to test-takers distributing their time on items differently. Furthermore, we believe that this mechanism could (partly) explain the mixed findings regarding item order effects in the literature. In the following section, we illustrate how speededness can be defined and how likely it is to occur. For this, we first introduce a modeling framework that allows us to quantify ability and speed as latent constructs.

Modeling Framework

To investigate the effects of item ordering and speededness on test performance, a joint model for speed and ability is required. Note that we model responses and response times at the level of the item, and not at the level of the complete test. A convenient choice for the item response model is the Rasch model (Rasch, 1960). The Rasch model

assumes that the probability of giving a correct response depends on a person parameter θ_i $i = 1, ..., n$, representing the person's ability (*ability parameter*) and an item parameter b_k $k = 1, ..., j$, representing the difficulty of the item (*difficulty parameter*):

$$P(y_{ik} = 1|\theta_i, b_k) = \frac{\exp(\theta_i - b_k)}{1 + \exp(\theta_i - b_k)}. \quad (1)$$

A useful property of the model is the fact that sum scores per person or item are sufficient statistics for the ability and difficulty parameters, respectively. This means that the number of correctly answered items by a person can function as a proxy for ability (*number-correct scoring*, a scoring approach that is very common for university exams). We use the terms test score and ability estimate interchangeably in this paper.

The most common model for modeling response times in cognitive testing situations is the lognormal model by van der Linden (2006), which assumes that response times are lognormally distributed. The model can be written as

$$\ln RT_{ik} = \lambda_k - \zeta_i + \epsilon_{ik}, \quad \text{with} \quad \epsilon_{ik} \sim N(0, \sigma^2_{\epsilon_k}). \quad (2)$$

The *item time intensity* in the model is represented by λ_k, the *person speed parameter* is represented by ζ_i. Note that both parameters often have substantial correlations with their ability counterparts (item difficulty and person ability) but are indeed separate parameters. ϵ_{ik} represents an item and person specific residual which is normally distributed with mean 0 and an *item specific variance* $\sigma^2_{\epsilon_k}$. The joint hierarchical framework by van der Linden (2007) assumes joint multivariate normal item and person parameter distributions for the two dimensions ability and speed and allows the simultaneous estimation of both models.

Speededness

In his work, van der Linden (2011) formally defines test speededness as an interaction of the time limit of a test, the amount of work a test requires and the working

speed of the test-taker. This means a test is speeded for a test-taker if, given his/her optimal working speed, the person would run out of time before answering to all items. A useful concept for understanding test speededness is the so-called within-person speed-accuracy trade-off (Goldhammer, 2015). The trade-off refers to the fact that the *accuracy* or *effective ability* of a person (meaning the ability a person is able to show given a certain speed level) increases with increased amounts of time spent by the person on an item (see Figure 1). This increase has an upper bound: From a certain point on, additional time will not lead to more accurate answers. However, research in the area of response time modeling has shown that the speed distributions in test-taker samples are usually rather broad (van der Linden & Xiong, 2013), meaning that the working speed levels demonstrated by test-takers differ substantially. Meanwhile, practical constraints (e.g. limited space at universities) almost always require test administrators to use a fixed time limit in higher educational testing. Therefore, constructing unspeeded tests (so-called 'pure power tests') in the context of higher educational high-stakes testing is practically impossible (Goldhammer, 2015). Instead, most tests can be considered a mixture, where at least for a small proportion of the testing population a certain level of speededness occurs on the test.

When test-takers experience test speededness (i.e., they run out of time while working on a test), they are confronted with the following three options: (a) omit items, (b) increase working speed and decrease accuracy, and (c) not reach the end of the test. An extreme form of (b) would, for example, be rapid guessing. In the context of number-right scoring, (a) is seen to be less favorable than (b) or (c), because, for example, guessing is expected to be not very time consuming while it still substantially increases the probability of a higher score (Millman et al., 1965). In high-stakes assessments, indeed omission rates are rather low and decreasing with increasing test experience of test-takers (Gafni & Melamed, 1994). In practice, for test-takers with an initial slow working speed, often a mixture of (b) from a certain point (Bolt et al., 2002; Goegebeur et al., 2008) and (c)

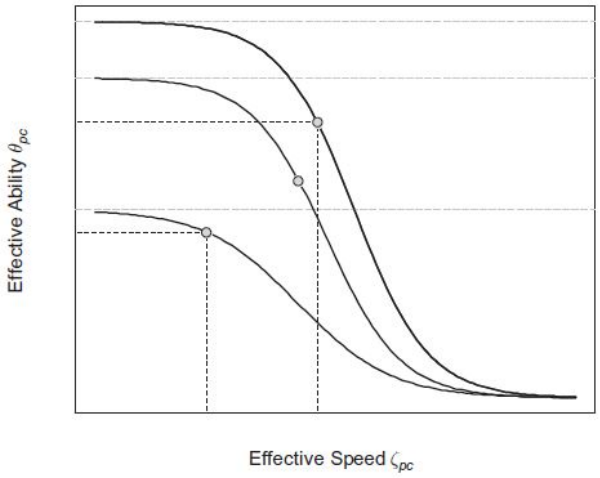

Figure 1
Speed accuracy trade-off as illustrated by Goldhammer (2015).

would be expected. As missing responses are usually scored incorrect in high-stakes assessments, all options (a) to (c) are reflected in lower test scores for test-takers that work under time pressure. Decisions of test-takers on which behavior to choose relate to the concept of test-wiseness.

Test-wiseness

Millman et al. (1965) define test-wiseness as "...a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score." They emphasize that the construct is usually logically independent of the actual measured construct. Therefore, it is commonly seen as a source of construct-irrelevant variance in the measured scores (Rogers & Yang, 1996). Furthermore, research has shown that test-wiseness is often unevenly distributed across subgroups, for example across different ethical backgrounds (Ellis & Ryan, 2003), and depends on the cultural match of the test-taker and the test (Melikyan et al., 2019). Therefore, test administrators often seek to minimize the influence of test-wiseness or specific test preparations on test scores, for example by giving clear instructions on the test or choosing item types less connected

to test-wiseness and test preparation (Powers, 1985; Powers & Rock, 1999).

A focal part of test-wiseness are time-using strategies (Millman et al., 1965). If a test is speeded for a test-taker, the test-taker has to allocate the available time in a way to maximize the expected score. This means that test-takers should identify and work on items that they are likely to answer correctly and for example guess on difficult items. Researchers have hypothesized that time-strategies might be culture-dependent, meaning the concept of speeded tests may be more prevalent in certain cultures than in others (Melikyan et al., 2020). It is apparent that the ordering of items determines the requirement for time-using strategies: For example, if items are sorted hard to easy, test-takers have to actively decide to spend less time on the initial items of the test. If items are sorted easy to hard, this decision is not required.

Consequences of different item orders under speededness

The introduced frameworks can be used to illustrate theoretical implications of different item orderings if a test is speeded: If a test-taker does not respond to all items at the end of a test or works with decreased accuracy, this negatively affects the person's test scores. How much the scores are affected, however, depends on the properties of the items that are not-reached or on which a higher speed was used, as already noted by Leary and Dorans (1985). Consider an example where a test-taker works linearly with a constant and insufficient working speed on a test with a fixed time limit (i.e., only option (c) occurs). In Table 1 such an example is illustrated, with not-reached items crossed out. The penalty for such a test taking behavior is much more severe on test form A, where three easy items are not-reached, than on test form B, where one hard item is not-reached. Note that such an effect is independent of specific item formats.

While it seems plausible to assume that most test-takers would speed up at the end of test form B in a realistic scenario, differences in the test scores would still occur between the two test forms. Obviously, it seems wisest for test-takers to distribute their time to

Table 1

Two Reversely Ordered Test Forms with Item Difficulties b_k and Expected Response Times for a Specific Speed Level ζ_i.

	Test form A		Test form B	
	b	RT	b	RT
Item 1	-0.1	10	1.2	60
Item 2	0.5	10	-0.6	30
Item 3	-0.2	20	~~-0.2~~	~~20~~
Item 4	-0.6	30	~~0.5~~	~~10~~
Item 5	~~1.2~~	~~60~~	~~-0.1~~	~~10~~

items in an adaptive fashion and to use (informed) guessing on difficult items (Dodeen, 2008; Millman et al., 1965). However, the test forms in Table 1 penalize lack of speed and time-using strategies very differently, namely: Both are much more important on test form B than on test form A. Note that test-wiseness also might vary strongly between assessment contexts: For some higher educational assessments, like TOEFL, GRE, ACT, or SAT test-takers and teachers sometimes spend considerable ressources on preparation, for example trying to increase test-takers' test-wiseness (Gulek, 2003; Kulik et al., 1984), with studies showing mixed findings but in general positive effects (Kulik et al., 1984). However, in the context of university exams, this may be less common.

The impact of different item orderings on test scores depends on the following factors: (a) the time limit of the test, (b) the working speed of the test-taker, (c) the time intensity of items at the end of the test, and (d) the difficulty of items at the end of the test. Factors (a) and (b) determine the general level of speededness of the test independent of the specific item ordering. Factor (c) determines the number of items the person will not reach or work with a decreased accuracy on, depending on the specific item ordering. Finally, factor (d) determines the impact of not-reached items or decreased accuracy at the end of the test.

Illustrative Example

To illustrate potential problems of different item orderings in speeded tests, we use data from an experimental administration of a high-stakes quantitative reasoning test. The data were collected as part of an study with various experimental conditions (Author et al., submitted). Participants were voluntary test-takers who wanted to prepare for the operational test and thus can be expected to be highly motivated. The overall correlation between the experimentally and later operationally measured ability was $r = .82$. The assessment contained 20 multiple-choice items. We analyze data from the conditions with a total time limit of 35 minutes. Feedback after every item was given to half of the students, but did not count towards the timing data. Item order was completely random for every test-taker. The data set consisted of 418 test-takers, of which 298 reported to be female and 119 reported to be male. The mean age in the sample was $M = 26.93$ ($SD = 5.97$). In total, 17 test takers were excluded from the analysis due to aborted test sessions (15 cases) and technical problems (2 cases).

Is the assessment speeded?

To investigate whether the assessment is speeded, we investigated number of not-reached items and performance decline coupled with speeding up at the end of the test. Skipping unanswered items was prevented within the assessment software. Of the 401 test-takers in the data set, 5.0% did not reach the end of the test (i.e. they ran out of time before answering to all items).

To investigate speeding up at the end of the test, we identified test-takers, who used almost all of the time available for the assessment. 127 of the 401 test-takers (31.7%) used more than 30 minutes. These test-takers are referred to as *slow test-takers*, whereas the other test takers are referred to as *fast test-takers*. In addition, for each test-taker, we split the test in two parts according to the item order: The first 15 items and the last five items. We compared the response accuracy and the response times in the first and last part using

proportion tests and median tests, respectively. Proportion correct were compared for the subset of slow test-takers including and excluding test-takers with not-reached items, as well as for fast test-takers. For both subsets of slow test-takers, on the first fifteen positions, the items are answered correctly more often (all slow test-takers: mean difference = 0.064, p = .009; slow test-takers without not reached items: mean difference = 0.065, p = .009). For fast test-takers, this difference is not meaningful (mean difference = 0.011, p = .517). The slow test-takers also take more time to answer to these items (median difference = 11.775, p = .001), while fast test-takers do not (median difference = 2.07, p = .10). Scatter plots with proportion correct and median response time on item level can be seen in Appendix Figures A1 and A2.

These findings indicate that for a substantial amount of test-takers the assessment was speeded. These test-takers performed better on the items at the beginning of the test than on the items at the end of the test. This was partially due to not-reached items but also due to taking less time on the items at the end of the test which resulted in decreased accuracy.

What are the potential consequences?

To illustrate the potential consequences of different item orderings, we simulated data for the slowest test-takers in the sample. First, we estimated a joint response and response time model. Based on the estimated parameters, we simulated responses and response times and implemented different item orderings. The goal was to compare differences in sum scores within the test-takers for different item orderings.

Data Simulation

We used the R package LNIRT (Fox et al., 2017) to estimate a joint hierarchical framework for responses and response times with the above described models. The estimated person and item parameters were used to simulate responses and response times for the seven slowest test-takers. Note that responses and response times were also

simulated for items that were originally not-reached for specific test-takers. We then applied different orderings of items to illustrate maximum potential bias between differently ordered test forms: (a) sorting items by increasing time intensity ('Short-Long'), (b) sorting items by decreasing time intensity ('Long-Short'), (c) sorting items by increasing difficulty ('Easy-Hard'), (d) sorting items by decreasing difficulty ('Hard-Easy'). These orderings were chosen to illustrate maximally unfair ordered test forms. Response times were then accumulated. If the cumulative response times exceeded the time limit of 35 minutes, the items were scored as incorrect (in a real exam, these items would have been not-reached). We then compared the resulting sum scores for the test-takers across the differently ordered test forms. Note that this approach simulates data with a constant working speed. In real life it seems plausible to assume that some test-takers would compensate running out of time by speeding up. However, as mentioned earlier, such behavior would also result in lower test scores due to decreased accuracy.

Results

Table 2

Simulated Test Scores for Different Item Orderings for Seven Different Test-Takers with Different Speed (ζ) and Ability Levels (θ) of one Randomly Chosen Replication.

ζ	θ	Short-Long	Long-Short	Easy-Hard	Hard-Easy	range(Σ)
-0.78	-0.87	5	1	5	2	4
-0.76	-0.83	3	3	3	3	0
-0.59	-0.02	7	5	8	4	4
-0.83	0.10	10	8	11	8	3
-0.64	-0.74	7	5	8	3	5
-0.59	-0.95	9	9	10	8	2
-0.58	-0.57	6	3	8	3	5

Note: Different item orderings means items were sorted in increasing or decreasing order by the respective item parameter time intensity (Short-Long or Long-Short) or difficulty (Easy-Hard or Hard-Easy). Columns contain the resulting test scores and column range(Σ) the maximum difference between these columns.

Table 2 illustrates that different item orderings can indeed lead to substantially different test results. For example, one of the most extreme results occurs for the person in

row five: On the test form with items sorted by increasing difficulty b the person achieves a sum score of 8, while on the test form with items sorted by decreasing difficulty b the person achieves a sum score of 3.

As the simulated responses and response times can vary substantially due to the probabilistic simulation process, we conducted 100 replications. The complete results can be seen in Appendix Table B1. For each of the seven test-takers the average range in sum scores between test forms across replications was greater than 2.5. The maximum difference across replications was between 6 and 10. These are substantial differences for a test with 20 items. Note that we chose the most extreme item orderings possible in this illustrative example. However, if different versions of a test are created by ordering items randomly, these extreme orderings are also possible.

Proposed Solutions

In this paper, we are proposing two solutions to avoid item position effects in higher educational assessments. First, a certain number of items at the end can be fixed in constant ordering across test forms. This prevents differential effects of item ordering at the end of a test, as test-takers run out of time on identical items. While some may argue that this reduces test security, we would argue that at the end of a test, test-takers are less likely to work on the same item compared to the beginning of the test, because test-takers work at different speed levels. Obviously it is not trivial to decide how many items or which portion of the test should have identical ordering at the end of the test forms. If too few items are chosen, test-takers might run out of time before the section is reached. If too many items are chosen, test security is lowered for no good reason. This can be seen as a security-fairness trade-off.

The effectiveness of the proposed approach can be enhanced by choosing to place the most time intensive items at the end of the test. By doing this, it becomes more unlikely that effects of speededness occur before the fixed set of items is reached. To

investigate the effectiveness of the proposed approaches we conducted a simulation study with realistic conditions for a higher educational exam.

Simulation Study

For the simulation study, hyper-parameters were used from the analyses of two psychology exams (organizational and social psychology) at a Dutch university. Hence, the simulated data is representative for the high-stakes higher educational testing context. Both exams contained 25 multiple-choice items and one open-answer item administered under a time limit of 40 minutes. The exams were conducted on computers in an online assessment setting and taken by 527 first-year psychology students. Students were not allowed to review items and all items were presented in a random order to the students. Responses and response times to all multiple-choice items were available for analysis, while item order was not. We analyzed the data using the R package LNIRT. As the results were very similar for both exams, we only report the results of the organizational psychology exam below. The hyper-parameters of the item and person parameter distributions are depicted in Appendix Table C1. The estimated correlation between item difficulty and time intensity was $r = 0.62$. The estimated correlation between speed and ability was $r = 0.24$.

Design

In the simulation, we used the illustrated hyper-parameters to create a realistic test containing 40 items. In each of the conditions, two test forms were created. We conducted the simulation study to answer the following questions: (a) Are the proposed approaches effective in preventing unfair effects of different item orders? (b) What are the effects if the number of items with fixed positions at the end of the test forms is too low? (c) What are the effects if time intensity is not known before the assessment and must be (imperfectly) predicted?

We varied two experimental factors: The number of items with fixed positions at the end of the test (three levels: [0; 5; 10]) and the selection and ordering of these items

(three levels: [random; based on an item time intensity covariate[2]; based on true item time intensity]). Because the second factor is irrelevant when the number of items with fixed positions is equal to zero, this resulted in overall seven conditions.

To observe a variety of speed and ability levels, person parameters were created as a grid: Speed levels were $[-0.6, -0.4, -0.2, 0]$ and ability levels were $[-1, 0, 1]$. These values were chosen because effects of speededness are relevant across all ability levels, but mainly relevant for slower test-takers. The grid also represents the width of possible person parameters according to Appendix Table C1. The time limit was set at 40 minutes. Responses and response times were created according to the Rasch model and the log-normal response time model (cf. above). Test scores were calculated. In total, 1000 replications were conducted. The complete R code for the simulation can be accessed here: https://osf.io/d97b5/?view_only=804fc3db7aab466e8cb358c6f7c7fa8c.

Results

To analyze unfairness of the test forms we compared test scores between the two test forms for all conditions and replications. In Table 3, the average and the maximum difference between the test scores on the test forms are depicted for all seven conditions. Note that the table only contains the results for the slowest but most able test-takers. The table illustrates that there can be considerable differences between two test forms with exactly the same items but different item orderings, if no measures are taken, with $M(\Delta) = 0.95$. The table can also be used to answer the research questions stated above: (a) Indeed, the proposed measures reduce differences between test forms. In the condition with the strongest control measures (ten items fixed, sorting based on time intensity), there are almost no differences between test forms on average, with $M(\Delta) = 0.19$. In fact, the simulation indicates that even imperfect measures serve the purpose of reducing effects of different item orderings, albeit less strongly. (b) If there are only five items fixed, which are

[2] Empirical mean correlation with true item time intensity of $r = .61$

sorted based on time intensity, the resulting mean difference between test forms is $M(\Delta) = 0.43$. (c) If the sorting occurs based on a covariate of time intensity, the resulting mean difference between test forms is $M(\Delta) = 0.22$. This indicates that number of items held constant is more important than quality of the time intensity prediction[3].

Table 3
Results of the Simulation Study: Mean (M(Δ)) and Maximum Difference (Max(Δ)) between the Test Scores for the Two Identical Test Forms with Different Item Orderings.

Fixed Positions	Ordering Items with Fixed Positions	M(Δ)	Max(Δ)
0	Random	0.90	4.08
5	Random	0.69	2.98
10	Random	0.32	1.36
5	Based on covariate	0.58	2.46
10	Based on covariate	0.22	1.07
5	Based on time intensity	0.43	1.90
10	Based on time intensity	0.19	0.68

Note: Item ordering was either completely random (0 items constant), or random with either the last five or ten items fixed. The constant items were either picked randomly or the most time intensive items ('Based on Time Intensity') or presumably most time intensive items were fixed ('Based on Covariate').

Complete results for all person parameter combinations can be seen in the Appendix Figures C1 and C2 for mean and maximum differences across replications, respectively. Results for the other person parameter combinations are comparable, albeit decreasing with increasing speed (test-takers run out of time less early) and decreasing ability (test-takers are less punished for not answering to items as they would have had a lesser chance of answering them correctly anyway).

Discussion

In the past, there have been various studies on whether different item orderings in higher educational testing are an adequate measure to increase test security or a potential

[3] Note that the maximum differences in sum scores between the test forms in Table 3 are less pronounced than in Appendix Table B1 despite the longer test forms. This is due to the fact that for Appendix Table B1 item orders have been specifically chosen to be as unfair as possible (e.g. Short-Long vs. Long-Short). Furthermore, in Table 3 results are aggregated across multiple test-takers while in Appendix Table B1 results are depicted for individual test takers.

source of unfairness. In a small illustration using quantitative reasoning data, we have shown that speededness plays a neglected but important role in the matter: When a test is speeded it becomes important to consider which items are placed at the end of a test, as these items are more likely to be not reached or test-takers allocate less time on them than on items at the beginning of the test. Furthermore, using the data set we illustrated how speededness can be detected by investigating missing responses and item position effects. To prevent such unfair test forms we proposed two straightforward measures to prevent effects of item ordering in speeded higher educational tests: Fixing the last items across test forms and additionally picking the most time intensive items for theses positions. In a simulation study based on data of Dutch university psychology exams, we illustrated that these approaches are indeed suitable to prevent unfair test forms regarding item ordering.

Practical Recommendations

From a practical point of view the question arises, how large the proportion with constant ordering at the end of a test should be and how time intensive items can be identified. In an ideal world, this should be determined be pretesting the test and determining the level of speededness. This could be done by using similar measures as in the illustrative example above or more complex modeling techniques such as change point analyses (Bolt et al., 2002; Goegebeur et al., 2008) However, a lot of higher educational exams and tests do neither have the opportunity to allow for extensive item pretesting without compromising test security nor have the required resources.

Hence, in many realistic settings, test administrators will have to rely on some assumptions and heuristics: Based on our analyses, we argue that holding one fourth of the items at the end of the test constant is a reasonable measure. Thereby test security is still not severely threatened but this proportion covers the part of the test on which changes to test taking behavior might be likely to occur. Note that the requirement of items held constant depends on the discrepancy between time intensities: If a single, very time

intensive item takes up one forth of the testing time for most test-takers, it might be sufficient just to put this single item at the end of the test. Moreover, if item difficulty and item time intensity are expected to be highly correlated, items that are anticipated to be difficult can be chosen for positions at the end of the test. Time intensity can also be expected to depend strongly on item type; open-answer items or elaborate constructed-response items can be expected to be almost always more time intensive than multiple-choice items. Finally, it should be noted that assigning time intensive items to the last item positions across test forms has the positive side effect of reducing the general influence of test-wiseness on test scores.

Alternative Approaches

Of course there are also different (but more complex) approaches to prevent problems regarding item order effects: Item time limits could be set to reduce differential effects of speededness (Goldhammer, 2015), as they prevent test-takers to distribute their time unwisely on the test. Furthermore, van der Linden and Xiong (2013) proposed a useful approach to control speededness in the framework of computer adaptive testing. While these approaches seem theoretically promising, they would often pose a substantial modification to higher educational assessment practice and require computer-based testing.

Alternatively, there is a wide range of psychometric models which aim at disentangling speed and ability. Even if effects of different item orderings have occurred, these models could be used to prevent bias in ability estimation (e.g., Pohl et al., 2019; Rose et al., 2017). However, most of these models were designed for use in low-stakes assessments and might be prone to gaming (e.g. test-takers purposely not reaching the end of the test). Furthermore, they require the availability of response times for analyses, which are only available in computer-based testing.

It is noteworthy that in some contexts test forms are created with no item overlap (e.g. different administrations of the GRE or TOEFL). In such situations, often approaches

known as automated test assembly are used to create parallel test forms (van der Linden, 2005). However, when such test forms are used, having exactly the same items fixed at the end of a test is impossible, as these test forms do not share the same items. Some of the mentioned approaches above (item time limits, CAT) may be able to solve the problem of unfair test forms due to different items at the end of test forms. However, the additional administration conditions that are required for these approaches may not be feasible in all testing situations where multiple test forms are assembled. Further research could investigate how fair test forms can be assembled in the context of test time limits and differences in speed between the test takers.

Conclusion

Although impact of item ordering on test fairness has been a topic of research for more than 50 year, the role of test speededness has been largely left unaddressed. In this paper, we have shown that especially when test-takers work under substantial time pressure and run out of time at the end of the test, item order plays a crucial role. Large differences in the expected test score are created between test forms that are supposed to be equivalent. To mitigate this issue, we have proposed two measures which keep the advantage of different item orders (increasing test security) while preventing unfair test forms: Keeping a certain number of, ideally time intensive, items constant at the end of a test. We believe that these measures can be easily implemented in practice and thereby help create fair test forms in the context of higher educational testing.

Ethics Declarations

The authors declare that the research presented in this manuscript is based on methodological considerations, simulated data and secondary data analysis. No human data was gathered for the presented research. To assure that simulated data was realistic, the parameters used to generate the data were based on aggregate statistics from an existing university exam. The collection of the data analyzed in the illustrative example

(the secondary data analysis) was approved by the ETS institutional research board. Participants were informed that the data was collected for research purposes only and that they could stop the test at any time without negative consequences.

References

Aamodt, M. G., & McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management*, *21*(2), 151–160.

Author, A., Author, B., & Author, C. (submitted). The impact of scoring instructions, time limits, and feedback on the measurement of quantitative reasoning. *Manuscript submitted for publication.*

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348.

Chen, H. (2012). The moderating effects of item order arranged by difficulty on the relationship between test anxiety and test performance. *Creative Education*, *3*(3), 328–333.

Chidomere, R. C. (1989). Test item arrangement and student performance in principles of marketing examination: A replication study. *Journal of Marketing Education*, *11*(3), 36–40.

Chirumamilla, A., Sindre, G., & Nguyen-Duc, A. (2020). Cheating in e-exams and paper exams: The perceptions of engineering students and teachers in Norway. *Assessment & Evaluation in Higher Education*, *45*(7), 940–957.

Davis, D. B. (2017). Exam question sequencing effects and context cues. *Teaching of Psychology*, *44*(3), 263–267.

Dodeen, H. (2008). Assessing test-taking strategies of university students: Developing a scale and estimating its psychometric indices. *Assessment & Evaluation in Higher Education*, *33*(4), 409–419.

Ellis, A. P., & Ryan, A. M. (2003). Race and cognitive-ability test performance: The mediating effects of test preparation, test-taking strategy use and self-eff icacy. *Journal of Applied Social Psychology*, *33*(12), 2607–2629.

Fox, J.-P., Klein Entink, R., & Klotzke, K. (2017). LNIRT: Lognormal response time item response theory models. R package version 1.995-0. https://cran.r-project.org/web/packages/LNIRT/index.html

Gafni, N., & Melamed, E. (1994). Differential tendencies to guess as a function of gender and lingual-cultural reference group. *Studies in Educational Evaluation*, *20*(3), 309–19.

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*(1), 65–87.

Goldhammer, F. (2015). Measuring ability, speed, or both? challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*, 133–164.

Gulek, C. (2003). Preparing for high-stakes testing. *Theory into Practice*, *42*(1), 42–50.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C.-l. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*(2), 179.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, *55*(3), 387–413.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

McKeachie, W. J., Pollie, D., & Speisman, J. (1955). Relieving anxiety in classroom examinations. *Journal of Abnormal and Social Psychology*, *50*, 93–98.

Melikyan, Z. A., Agranovich, A. V., & Puente, A. E. (2019). Fairness in psychological testing. In G. Goldstein, D. Allen, & J. DeLuca (Eds.), *Handbook of psychological assessment* (4th ed., pp. 551–572). Academic Press.

Melikyan, Z. A., Puente, A. E., & Agranovich, A. V. (2020). Cross-cultural comparison of rural healthy adults: Russian and american groups. *Archives of Clinical Neuropsychology*. https://doi.org/https://doi.org/10.1093/arclin/acz071

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25*(3), 707–726.

Monk, J. J., & Stallings, W. M. (1970). Effects of item order on test scores. *The Journal of Educational Research*, *63*(10), 463–465.

Neely, D. L., Springston, F. J., & McCann, S. J. H. (1994). Does item order affect performance on multiple-choice exams? *Teaching of Psychology*, *21*(1), 44–45.

Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology*, *39*(4), 299–307.

Pettit, K. L., Baker, K. G., & Davis, L. D. (1986). Unconscious biasing of student examination scores: A case of sequential versus random information retrieval. *Journal of Marketing Education*, *8*(3), 20–24.

Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response times to model not-reached items due to time limits. *Psychometrika*, *84*(3), 892–920.

Powers, D. E. (1985). Effects of coaching on GRE aptitude test scores. *Journal of Educational Measurement*, *22*(2), 121–136.

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement*, *36*(2), 93–118.

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Rogers, W. T., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment*, *12*(3), 247–259.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in irt models. *Psychometrika*, *82*(3), 795–819.

Russell, M., Fischer, M. J., Fischer, C. M., & Premo, K. (2003). Exam question sequencing effects on marketing and management sciences student performance. *Journal of Marketing Education*, *3*, 1–10.

Sax, G., & Cromack, T. R. (1966). The effects of various forms of item arrangements on test performance. *Journal of Educational Measurement*, *3*(4), 309–311.

Togo, D. F. (2002). Topical sequencing of questions and advance organizers impacting on students' examination performance. *Accounting Education*, *11*(3), 203–216.

van der Linden, W. J. (2005). *Linear models for optimal test assembly*. Springer.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, *48*(1), 44–60.

van der Linden, W. J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 418–438.

Vander Schee, B. A. (2013). Test item order, level of difficulty, and student performance in marketing education. *Journal of Education for Business*, *88*(1), 36–42.

Wang, L. (2019). Does rearranging multiple choice item response options affect item and test performance. *ETS Research Report Series*, *1*, 1–14.