Personality and Individual Differences 207 (2023) 112157

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Which tests should be administered first, ability or non-ability? The effect of test order on careless responding

Hanif Akhtar [a,b,*], Kristof Kovacs [c]

[a] Doctoral School of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
[b] Faculty of Psychology, Universitas Muhammadiyah Malang, Malang, Indonesia
[c] Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary

ARTICLE INFO

Keywords:
Test-taking effort
Ability test
Non-ability test
Web-based survey
Response time effort

ABSTRACT

Responses provided by unmotivated survey participants can threaten both the quality and validity in psychological assessment. When two types of tests (ability and non-ability) are administered, it is not obvious which test should be administered first in order to maximize effort in responding by participants. Currently, there is no applicable guideline for researchers. We aimed to examine whether the order of presentation affects effort in responding. We tested university students (N = 367) in an online experimental study. Participants were randomly assigned to one of two conditions: (a) ability tests are administered first, and (b) non-ability tests are administered first. The results indicate that test order does influence participants' test-taking efforts. Taking non-ability tests first resulted in significantly higher effort for non-ability tests. For ability tests, the order of presentation did not matter: neither effort nor performance varied as the function of the order of presentation in the case of ability tests. At the same time, participants' test-taking effort was negatively correlated with item and scale position. Based on our findings, we suggest administering non-ability tests first. This suggestion should be limited to the context of unproctored online low-stakes surveys.

1. Introduction

Research in psychology frequently relies on the motivation of research participants to provide their maximum performance on low-stakes assessments (e.g., surveys). When there are no personal consequences for participants, it is reasonable to assume that some participants will answer carelessly. Unmotivated response behaviour can occur in various low-stakes contexts, including surveys, educational assessments, and training evaluations. Many terms have been used in the literature to refer to the phenomenon in which participants are unmotivated to complete a survey and genuinely trying to arrive at a correct solution (see Arthur et al., 2021; Huang et al., 2012). For example, many terms used were random responding (Pinsoneault, 2007), careless responding (Meade & Craig, 2012), protocol invalidity (Johnson, 2005), and insufficient effort responding (Huang et al., 2012). We use the term *careless responding* to collectively refer to these terms. We define careless responding as non-systematic responses with deliberate disregard for item content due to a lack of test-taking effort. The terms careless responding and (a lack of) test-taking effort will be used interchangeably in this paper. Meade and Craig (2012) claim that careless responding is likely to be an issue when data is gathered through anonymous web-based surveys, especially with student samples.

The detection and consequences of careless responding are well-documented (for a review, see Arthur et al., 2021; Huang et al., 2012). Careless responding could be problematic because it results in noisy data which, in turn, leads to biased measurement properties (Wise & DeMars, 2006). This has several negative consequences, including attenuating the observed internal structure of measures and weakening correlations (Credé, 2010; DeSimone & Harms, 2018; Johnson, 2005). Specifically, in a research setting, careless responding might lead to incorrect conclusions being drawn from the results (Ophir et al., 2020). Therefore, the concern about careless responses in research settings is reasonable.

Instruments in research may be classified according to whether they are tests of ability/achievement (hf. ability tests) or personality/attitude/etc. (hf. non-ability tests): on the former, one can provide correct and incorrect answers, whereas, on the latter, one cannot. In a high-stakes assessment context, cheating is the main concern in ability testing, and dishonest responses in non-ability testing (Arthur et al., 2021). However, low-stakes assessments (including surveys) have a

* Corresponding author at: Doctoral School of Psychology, ELTE Eötvös Loránd University, Izabella utca 46, 1064 Budapest, Hungary.
E-mail address: akhtar.hanif@ppk.elte.hu (H. Akhtar).

https://doi.org/10.1016/j.paid.2023.112157
Received 7 July 2022; Received in revised form 10 February 2023; Accepted 4 March 2023

similar threat to the validity of the score: careless responding.

In the online survey context, participants typically need to invest more effort in answering items on ability tests than on non-ability tests. Non-ability tests typically use items with similar content and format repetitively. As a result, even initially motivated participants may lose interest and become bored as they progress through the test, increasing their likelihood of responding carelessly (Galesic, 2006). Previous research on non-ability tests discovered a positive correlation between measure length and careless responding (Gibson & Bowling, 2020). In ability tests, previous research found that participants show a decrease in motivation if the test is challenging (Barry & Finney, 2016). Item difficulty (Asseburg & Frey, 2013), item location (Pastor et al., 2019), and item type (Sundre & Kitsantas, 2004) all have an impact on test-taker motivation. Thus, the causes of participants' careless responses may be different for ability and non-ability tests.

Previous studies suggested that test-taking motivation can fluctuate during testing sessions (Barry et al., 2010; Barry & Finney, 2016; Pastor et al., 2019; Penk & Richter, 2017). Wolgast et al. (2020) found that test order affected test-taker effort. They found intraindividual differences in students' efforts; specifically, students' efforts did not decrease when the cognitive ability test came first, and a mock exam came second but significantly decreased when the mock exam came first, and the ability test followed. The author's explanation for these findings was that students had a higher level of accuracy on the cognitive ability test than on the mock exam. They also found that presenting an easier test at the beginning of the testing session could be more motivating. However, they only included ability tests in their study.

Other studies examined changes in test-taking efforts on classroom assessment using one ability test and four non-ability tests (Barry et al., 2010; Barry & Finney, 2016). They found that students reported more effort on the non-ability tests and less effort on the ability test. These results demonstrated that students are less motivated to exercise in low-stakes assessments. Barry and Finney (2016) also found that effort slightly changed during the testing session: the ability test that was administered first was the least motivating. This suggests that in low-stakes assessments, the order of the tests matters. However, the order of the test was not directly manipulated in the study. In addition, effort was only measured with self-report, and their findings were limited to paper-and-pencil tests in a classroom assessment context. In the context of online research, when both ability and non-ability tests are administered, their order might be important.

In some areas of research as well as in certain low-stakes assessments for international comparisons (e.g., PISA), ability and non-ability tests are administered together. In such instances, which kind of test should be presented first in order to maximize participants' effort? There is no applicable guideline for researchers that addresses this question, and empirical evidence regarding the effect of test order on careless responding is limited.

1.1. Current study

Previous studies have been limited in providing practical recommendations to researchers on test order that could maximize effort in responding by research participants. Specifically, there are three research questions (RQ) in our study.

RQ1. Are there any differences in the responding behaviours of participants as the function of which test is administered first?
RQ2. Does item and scale position correlate with careless responding?
RQ3. How is the reliability of the respective measures affected by presentation order?

These questions are examined in the context of a low-stakes web-based survey conducted on university students. The response time of participants was recorded for each item and scale. We randomly manipulated the order of the test presentation. Two indicators of careless responding were used: self-reported effort (SRE) and response time effort (RTE) (Wise & Kong, 2005). SRE provides a global measure of test-taking effort based on participants' self-ratings right after completing tests. RTE is a time-based measure based on the hypothesis that unmotivated participants will answer too quickly (i.e., before they have time to read and fully consider it) when administered an item or scale.

2. Methods

2.1. Participants

A total of 589 students (429 females, 72.8 %) participated in this study. The participants, aged 18 to 46 ($M = 21.6$, $SD = 3.79$), were all Indonesian university students; 529 were undergraduates, 43 were master's students, and 17 were doctoral students. An online survey was used to collect data. Participants were recruited in April 2022 through advertisements on social media (Facebook, Instagram, and WhatsApp groups) and flyers. Thirty participants selected randomly received a monetary reward of 50,000 Indonesian Rupiah each (3.5 Euros). The respondents participated voluntarily and consented to their anonymously collected data being analyzed. Participants were asked to complete ability and non-ability tests in a different, randomized order. Informed consent was obtained from all individual participants involved in the study. All scales used in this study were programmed with the PsyToolkit platform (Stoet, 2016).

2.2. Measures

The ability tests used in this study were the 16-item International Cognitive Ability Resource sample test (ICAR-16) and the Cognitive Reflection Test (CRT). The non-ability tests used in this study were the Ten Item Personality Inventory (TIPI), Self-Estimated Intelligence (SEI), Self-Reported Cognitive Abilities Questionnaire (SRCAQ), and the Multidimensional Competitive Orientation Inventory (MCOI). Finally, participants' careless responses were measured using two indicators: effort thermometer scores and RTE (further details in the supplements).

2.2.1. ICAR-16

ICAR-16 is a 16-item cognitive ability test from The International Cognitive Ability Resource project (Condon & Revelle, 2014). A 16-item test consists of four item types, i.e., verbal reasoning, letter and number series, matrix reasoning, and three-dimensional rotation. The items used a multiple-choice format.

2.2.2. CRT

CRT (Frederick, 2005) is a brief measure of cognitive ability. CRT measures cognitive processing, especially the tendency to suppress an incorrect answer that comes to mind intuitively and comes at a more deliberate correct answer. The CRT consists of three mathematical tasks with a short-answer response format.

2.2.3. TIPI

TIPI (Gosling et al., 2003) is a short instrument measuring the Big-Five dimensions (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness). TIPI consists of 10 items (e.g., “Anxious, easily upset”). Participants responded on a 7-point Likert scale, ranging from 1 (disagree strongly) to 7 (agree strongly).

2.2.4. SEI

SEI was designed to measure global self-estimated intelligence (Furnham, 2001). In this part, participants were provided with an image of a normal distribution with standard IQ scores along with descriptions of IQ. Participants were asked to provide estimates between 55 and 145 for their IQ.

2.2.5. SRCAQ

SRCAQ was a brief measure designed to self-assess cognitive functioning in fluid reasoning, short-term working memory, long-term storage and retrieval, comprehension knowledge, processing speed, auditory processing, and visual processing. The SRCAQ consists of 21 items (e.g., "Imagine what an object looks like when rotated or reflected"). Participants responded on a 5-point Likert scale, ranging from 1 (very difficult) to 5 (very easy).

2.2.6. MCOI

The MCOI (Orosz et al., 2018) is a multidimensional scale measuring competitive orientation. MCOI consists of 15 items (e.g., "The most important thing is winning, no matter what"). Participants responded on a 6-point Likert scale, ranging from 1 (not like me at all) to 6 (completely like me).

2.2.7. Effort thermometer

The effort thermometer (Baumert & Demmrich, 2001) is a self-report measure designed to assess the investment of effort on an anchored scale. Participants were asked to imagine a situation that would be extremely important to them personally and in which they would invest maximum effort (e.g., a university entrance exam). This maximum effort was assigned a value of 10. Participants were then asked to rate how much effort they had just invested in the test in relation to the maximum-effort situation on a scale ranging from I to 10. This measure is used in the PISA.

2.2.8. RTE

RTE is a time-based measure of effort. It is based on the hypothesis that unmotivated participants will respond too quickly when administered an item or scale. We used the 30 % Normative Threshold (NT30) approach (Wise & Kuhfeld, 2020) because it may be preferred for research purposes (Soland et al., 2021), and it is more appropriate in the context of non-ability tests (Johnston, 2016). This approach proposes that 30 % of the average response time be used for the threshold. For example, if it takes participants an average of 50 s to respond to an item, NT30 would be 15 s. In case a participant responds quicker than this threshold, their response is considered rapid-guessing behaviour (RG), qualifying as careless responding. In contrast, if a participant responds slower than the threshold, their response is considered appropriate solution behaviour (SB). Participants j, with response times to item i at or above the threshold were given a 1 on the SB index, and 0 otherwise.

The index of overall RTE for participant j is calculated by summing the SB index values across all items and dividing by the number of items in the test (k).

$$RTE_j = \frac{\sum SBij}{k}$$

For non-ability tests, RTE was calculated based on response time at the scale level instead of the item level. Thus, RTE for participant j is calculated by summing the SB index values across all scales and dividing by the number of scales. However, the interpretation of RTE for item-level or scale-level analysis is still the same: the proportion of solution behaviour during the testing session. RTE values near 1 indicate maximal effort, and values near 0 indicate minimal effort.

2.3. Procedure

After completing a demographic survey, participants were randomly assigned to two conditions. In group 1, participants responded to the ability tests first and the non-ability test second, while the order was the opposite for group 2. The total number of items for ability tests was 19. The items of ICAR and CRT were mixed. Items were presented one by one in the same order for each participant. The total number of items of non-ability tests was 47. The scales were presented sequentially in the following order: TIPI, SEI, SRCAQ, and MCOI. Items in each scale were presented on the screen at once.

The time needed to answer each page of items was recorded (item level for ability tests and scale level for non-ability tests). After completing one set of tests (ability or non-ability), participants were asked to indicate on a scale ranging from 1 to 10 how much effort they had just invested in completing the tests. There was no time limit, and most participants completed the entire study in <60 min. Fig. 1 shows the study design. The study was approved by the ethical committee at the authors' university (Number 2022/151-2).

2.4. Analyses

To answer RQ1, we used 2 × 2 Mixed ANOVA. After completing ability and non-ability tests, participants' effort (T1 and T2) was used as a repeated measure factor, and the experimental group (group 1 or group 2) was used as between-subject factors. Analyses were conducted separately for self-reported effort and RTE. Multiple regression analyses were used to examine whether personality and cognitive abilities influenced test order effects.

To answer RQ2, we used Spearman's rank order correlation. Analysis was conducted separately for ability and non-ability tests. Participants' effort (measured by average SB index) was correlated with item and scale position. Visual inspection was used to get additional insight into the effect of the item or scale position. Finally, to answer RQ3, we compared the alpha coefficient of each measure in each group. We used Feldt's 95 % confidence interval to test whether the difference was significant.

3. Results

3.1. Preliminary analysis

Incomplete answers were removed from the analysis; 222 participants (37.7 %) did not complete the entire survey. The final dataset contained 367 participants. The dropout rate was 42.9 % and 32.3 % for groups 1 and 2. The average self-reported effort scores before participants left were 6.88 and 8.64 for the ability and non-ability tests, respectively (further details in the supplements). The incomplete data might indicate a low motivation to complete the survey. However, technical issues (e.g., bad internet connection) might have also caused participants to drop out. A separate analysis using imputed datasets was performed on fully randomized samples. Please find supplementary materials for details.

The number of participants in each group, mean, standard deviation, and correlation coefficients among variables are presented in Table 1. The two measures of effort (self-report and RTE) were positively correlated. The correlation coefficient between RTE and self-report for ability tests was $r = 0.32$, and for non-ability tests was $r = 0.36$. There was no difference between groups 1 and 2 in effort or performance when completing ability tests (ICAR-16). However, group 2 exerted higher effort when completing non-ability tests.

3.2. Test-taking effort across conditions

A 2 × 2 ANOVA was performed. The analysis of RTE revealed a significant main effect of effort measures (T1 and T2) ($F(1, 365) = 15.0$, $p < .001$), and interaction between effort measure and group membership ($F(1, 365) = 46.4$, $p < .001$). However, the main effect of group membership was non-significant ($F(1, 365) = 1.02$, $p > .05$). The same conclusion was found for self-reported effort. The main effect of effort measures ($F(1, 365) = 9.81$, $p < .001$), and the interaction between effort measure and group membership ($F(1, 365) = 71.8$, $p < .001$) were significant, but a significant main effect of the experimental group was found ($F(1, 365) = 4.03$, $p < .05$). Mixed ANOVA results are presented in Table 2.

Fig. 2 shows marginal means plots among conditions. The pattern

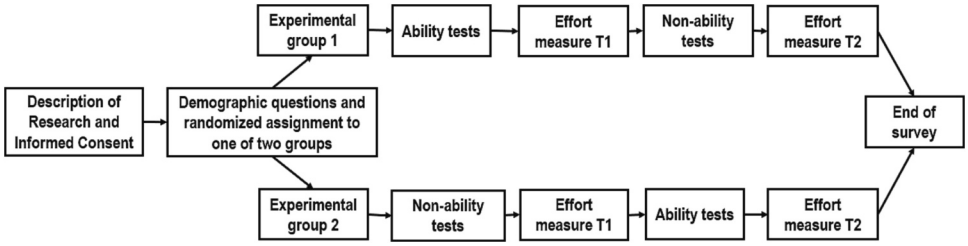

Fig. 1. The study design.

Table 1
Descriptive statistics and correlation coefficients.

Variables	Group 1 (ability 1st)		Group 2 (non-ability 1st)		*d*	1	2	3	4	5
	N	Mean (SD)	N	Mean (SD)						
1. RTE Ability	172	0.87 (0.20)	195	0.85 (0.21)	0.11	–				
2. RTE Non-ability	172	0.91 (0.23)	195	0.96 (0.14)	0.31**	0.46**	–			
3. SRE Ability	172	7.75 (2.08)	195	7.74 (2.06)	0.03	0.32**	0.15**	–		
4. SRE Non-ability	172	8.31 (1.80)	195	8.97 (1.39)	0.40**	0.28**	0.36**	0.41**	–	
5. ICAR-16	172	0.46 (0.23)	195	0.45 (0.20)	0.07	0.48**	0.13*	0.28**	0.18**	–
6. CRT	172	0.24 (0.32)	195	0.18 (0.27)	0.21*	0.17**	0.08	0.08	0.06	0.49**

Note: SRE = self-reported effort, RTE = response time effort, ICAR-16 = 16-item cognitive ability test from The International Cognitive Ability Resource, CRT = Cognitive Reflection Test.
* $p < .05$.
** $p < .01$.

Table 2
Mixed ANOVA results of Participants' effort.

	Response time effort					Self-reported effort				
	df	Mean square	F	p	$g\eta^2$	df	Mean square	F	p	$g\eta^2$
Within subjects effects										
Effort	1, 365	0.02	15.0	<.001	0.011	1, 365	4.78	9.81	<.001	0.008
Effort * Group	1, 365	0.02	46.4	<.001	0.032	1, 365	2.04	71.8	<.001	0.056
Between subjects effects										
Group	1, 365	0.06	1.02	.310	0.002	1, 365	2.04	4.03	.046	0.008

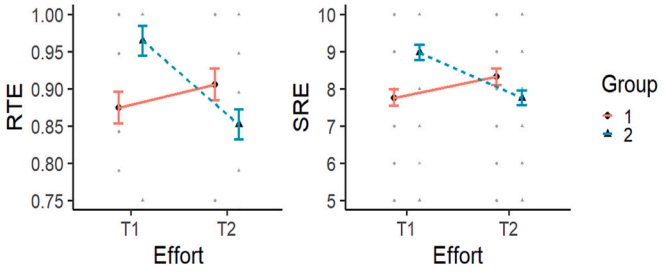

Fig. 2. The interaction of experimental group and effort measures in T1 and T2.
Note: RTE = response time effort, SRE = self-reported effort.

was similar for RTE (left) and self-reported effort (right). Overall, participants' effort to complete non-ability tests was higher than their effort to complete ability tests. There was no difference in the participants' effort to complete the ability tests as the function of test order. However, this was not the case for non-ability tests: when non-ability tests were administered first (group 2), participants' effort was higher than when ability tests were administered first (group 1).

Multiple regression analyses were performed to examine whether individual characteristics influenced effort (further details in the supplements). After controlling cognitive abilities and personality traits, group membership was a significant predictor of test-taking effort for the non-ability tests but not for ability tests. This finding was consistent with previous analysis.

3.3. Test-taking effort across item and scale position

Spearman rank-order correlation was performed. Item position significantly correlated with participants' effort ($r = -0.59$, $p < .05$), even after controlling for item difficulty (proportion of correct answers) ($r = -0.57$, $p < .05$). For non-ability tests, since the sample size was only four, the statistical analysis might be meaningless. Fig. 3 shows the effect of item and scale position on participants' efforts. Generally, there is a decreasing trend in participants' efforts as the test progresses.

3.4. The effect of presentation order on reliability

We compared the alpha coefficient of each measure in each group to examine whether the order presentation affects the measurement properties (detailed results in the supplements). In general, there was a trend that instruments presented first had higher alpha than those presented second, except for SRCAQ. However, the difference was not significant in any of the comparisons. Therefore, we can conclude that in this study, we did not demonstrate an effect of the order of presentation on internal consistency reliability.

4. Discussion

The main goal of this study was to examine whether the order of test

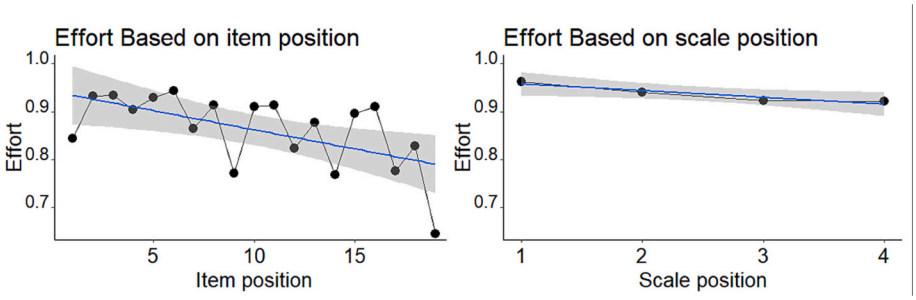

Fig. 3. Correlation between items and scale position and participants' effort to complete the tests.

presentation affects participants' careless responses. Our analyses suggested that test order does influence participants' efforts: when non-ability tests were administered first, participants' effort was higher for non-ability tests than when ability tests were administered first. On the other hand, the order of presentation did not matter for ability tests. However, many participants reported high effort during the session, even though the survey was low-stakes.

Similarly, an analysis of dropout participants showed that presenting the ability test first is detrimental to participants' motivation. Most participants who did not complete the survey were administered ability tests first. The average SRE scores for ability tests before participants left were lower than those of participants who completed the survey. A previous study found that participants who dropped out after a particular block reported significantly worse subjective experiences with that block (Galesic, 2006). Assuming that the participants dropped out because of low motivation, taking mentally exhausting ability tests at the beginning of the survey could have played a role.

Overall, participants' effort to complete non-ability tests was higher than their effort to complete ability tests. This finding is consistent with previous findings (Barry et al., 2010; Barry & Finney, 2016). From the expectancy-value theory (Wigfield & Eccles, 2000) perspective, the mental taxation required to answer the ability test items correctly could lead to low expectancy, resulting in a low effort. In contrast to the less demanding non-ability tests, ability tests were cognitively demanding. In accordance with previous studies (Pastor et al., 2019; Wise & Kingsbury, 2016), we also found that participants' effort was negatively correlated with item and scale position. Thus, our findings confirm the effect of item position on participants' efforts.

We found a trend that the instrument administered first had somewhat higher reliability, although the difference is not significant. The exception is for SRCAQ. The characteristics of the scale could explain this finding. SRCAQ is a scale with no reverse-coded items. Participants who answer carelessly might provide consecutive identical responses, which will artificially increase the alpha coefficient (see supplements for details). Therefore, interpreting the high-reliability coefficients of the all-positive keyed scale should be cautious, as invalid data responses might contaminate it.

4.1. Implications for research design

We found that administering non-ability tests prior to ability tests could be beneficial in reducing participants' careless responses. On the other hand, we found that participants' efforts to complete ability tests did not differ as the function of the order of presentation. In addition, we did not find any difference in performance on the ability test either, i.e. earlier or later presentations did not have an effect on the ability estimates. Based on our findings, the simple suggestion for designing data collection is to administer non-ability tests first. This suggestion should be limited to the context of unproctored online low-stakes surveys with similar test types and testing conditions as our study.

Our findings also have implications for data collection with the purpose of estimating item parameters, such as item difficulty under Rasch measurement or item-response theory. Items at the end of the test may be answered randomly by the participants, which results in biased parameters. Therefore, the randomization of items to be presented for calibration purposes is crucial.

4.2. Limitations and directions for future research

First and foremost, our sample is limited to university students. Other participants might have different motivations for participating in a survey. Second, gender was not equally distributed in our study, with twice the number of females. As males were found to engage in rapid guessing nearly twice as often as females (Soland, 2018), this difference might matter. Third, the SB index for non-ability tests was calculated on the scale level because response time was recorded for each page only. Further investigation using item-level analysis is needed to confirm our results. Fourth, we measured effort after administering a series of instruments (set of ability or non-ability tests). However, participants' efforts might be unique for a specific test. Further investigation acknowledging the instrument-specific effects of the individual instruments is needed. Fifth, we only used two indicators of careless responding as those were the only indicators suitable for our study design. However, other indices of careless responding are available (see Arthur et al., 2021) and can be used for further studies.

Finally, several confounders, such as participants' fatigue, might affect the result of the experiment. As subjective fatigue increases with increasing test length (Ackerman & Kanfer, 2009), future research can replicate this study by manipulating the length of the test. Namely, if shorter ability tests are administered first, is it still detrimental to participants' efforts? Otherwise, will a negative effect occur if longer non-ability tests are administered first? Other test characteristics might also influence test-taking efforts, such as item difficulty, order of item difficulty, item type, and response type. Replication by manipulating those features is needed to test the generalizability of our findings.

5. Conclusion

The common practice in high-stakes assessments when two types of tests are presented (e.g., in personnel selection) is to administer ability tests first. Ability tests are more mentally exhausting than non-ability tests. By presenting an ability test at the beginning, the test-taker is expected to provide maximum performance. However, test-taking effort is not of great importance in high-stakes assessments since test-takers are assumed to give maximum effort, regardless of the order of presentation. In a low-stakes assessments context, however, test-taking effort is not necessarily high. Our study shows that, for low-stakes assessment contexts, presenting ability tests first is detrimental to participants' test-taking efforts. The conclusion should be limited to similar test types and testing conditions.

In sum, we recommend presenting non-ability tests first. This recommendation should be limited to the context of unproctored web-based low-stakes surveys. This study also confirms that item and scale position affects participants' efforts. However, replication with different

measures, sample characteristics, data collection protocols, and analysis procedures should be performed in order to test the generalizability of our findings.

Funding

Kristof Kovacs received funding by the National Research, Development and Innovation Office of Hungary: Grant FK-21-138971, by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and by the ÚNKP-22-5 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund.

CRediT authorship contribution statement

Hanif Akhtar: Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Writing – original draft. **Kristof Kovacs:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition.

Data availability

The datasets analyzed can be found in the OSF Repository (https://osf.io/6b7zm/?view_only=95d89cecbfc14d719f1343e6cd9ea382)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.paid.2023.112157.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15*, 163–181. https://doi.org/10.1037/a0015719

Arthur, W., Hagen, E., & George, F. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior, 8*, 105–137. https://doi.org/10.1146/annurev-orgpsych-012420-055324

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*(1), 92–104.

Barry, C. L., & Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education, 29*(1), 46–64. https://doi.org/10.1080/08957347.2015.1102914

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342–363. https://doi.org/10.1080/15305058.2010.508569

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441–462. https://doi.org/10.1007/BF03173192

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence, 43*(1), 52–64. https://doi.org/10.1016/j.intell.2014.01.004

Credé, M. (2010). In *, 70. Random Responding as a Threat to the Validity of Effect Size Estimates in Correlational Research* (pp. 596–612). https://doi.org/10.1177/0013164410366686 (4).

DeSimone, J. A., & Harms, P. D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology, 33*(5), 559–577. https://doi.org/10.1007/S10869-017-9514-9

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Furnham, A. (2001). Self-estimates of intelligence: Culture and gender difference in self and other estimates of both general (g) and multiple intelligences. *Personality and Individual Differences, 31*(8), 1381–1405. https://doi.org/10.1016/S0191-8869(00)00232-4

Galesic, M. (2006). Dropouts on the web: Influence of changes in respondents' interest and perceived burden during the web survey. *Journal of Official Statistics, 22*(2), 313–328.

Gibson, A. M., & Bowling, N. A. (2020). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment, 36*(2), 410–420. https://doi.org/10.1027/1015-5759/A000526

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality, 37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*(1 SPEC. ISS.), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Johnston, M. M. (2016). *Applying solution behavior thresholds to a noncognitive measure to identify rapid responders: An empirical investigation.* James Madison University.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/A0028085

Ophir, Y., Sisso, I., Asterhan, C. S. C., Tikochinski, R., & Reichart, R. (2020). The turker blues: Hidden factors behind increased depression rates among Amazon's mechanical turkers. *Clinical Psychological Science, 8*(1), 65–83. https://doi.org/10.1177/2167702619865973

Orosz, G., Tóth-Király, I., Büki, N., Ivaskevics, K., Bothe, B., & Fülöp, M. (2018). The four faces of competition: The development of the multidimensional competitive orientation inventory. *Frontiers in Psychology, 9*(MAY). https://doi.org/10.3389/FPSYG.2018.00779

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment, 24*(3), 189–212. https://doi.org/10.1080/10627197.2019.1615373

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability, 29*(1), 55–79. https://doi.org/10.1007/s11092-016-9248-7

Pinsoneault, T. B. (2007). Detecting random, partially random, and nonrandom Minnesota multiphasic personality inventory-2 protocols. *Psychological Assessment, 19*(1), 159–164. https://doi.org/10.1037/1040-3590.19.1.159

Soland, J. (2018). Are achievement gap estimates biased by differential student test effort? Putting an important policy metric to the test. *Teachers College Record, 121*(12), 1–26. https://doi.org/10.1177/016146811812001202

Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education, 9*(1). https://doi.org/10.1186/s40536-021-00100-w

Stoet, G. (2016). In *, 44. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments* (pp. 24–31). https://doi.org/10.1177/0098628316677643, 1.

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-talking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*(1), 6–26. https://doi.org/10.1016/S0361-476X(02)00063-2

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement, 53*(1), 86–105. https://doi.org/10.1111/jedm.12102

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Kuhfeld, M. R. (2020). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement, 58*(1), 130–149. https://doi.org/10.1111/jedm.12275

Wolgast, A., Schmidt, N., & Ranger, J. (2020). Test-taking motivation in education students: Task battery order affected within-test-taker effort and importance. *Frontiers in Psychology, 11*. https://doi.org/10.3389/FPSYG.2020.559683