The Spanish Journal of Psychology (2026), **29**, e8, 1–10
doi:10.1017/SJP.2026.10022

Article

Use of Crossed Random-Effects Models to Assess Multiple-Choice Items: An Experimental Study

Susana Sanz[1,2], Carmen García[2] and Ricardo Olmos[2]

[1]Department of Psychology, Universidad Pontificia Comillas, Spain and [2]Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Spain

Abstract

Including general options on items, such as None-of-the-above (NOTA), could lead to worse psychometric properties. Further, personality traits are related to academic performance and could influence NOTA effects. Therefore, this study aims to test the effects of NOTA by manipulating its appearance and its use as the correct option or as a distractor, while considering the examinees' personality traits. The sample consisted of 449 psychology students who answered a statistics concept inventory. A crossed random-effects model was conducted to model the probability of answering an item correctly depending on items and students' covariates. The appearance of NOTA negatively affects the probability of answering correctly and changes between examinees. Conscientiousness, Extraversion, Emotional Stability, and previous knowledge of statistics are also significant predictors. The results point to increased items' difficulty when NOTA is included, although not all the examinees seem to show the same behavior when dealing with NOTA options.

Keywords: crossed random-effects models; educational assessment; multiple-choice items; none-of-the-above; personality

(Received: 27 February 2025; revised: 15 October 2025; accepted: 19 February 2026)

Introduction

Measuring the students' knowledge in a particular subject is not easy. Often, months or even years of study must be summarized in a single assessment. For this reason, it is important to write proper exams within the academy, which involves time and effort for lecturers. In this regard, multiple-choice items (MC) are popular nowadays for measuring knowledge. They allow evaluating a considerable number of students in an accurate, objective, and quick way that makes lecturers able to sample efficiently large syllabus (Haladyna et al., 2019).

For an exam to be effective, items' difficulties should vary, predominating the items with an intermediate difficulty. More importantly, items must possess strong psychometric properties to effectively differentiate between examinees' knowledge levels. That is why the stems and response options must be clear enough and avoid error induced by poor writing so only the examinees' knowledge is assessed. Additionally, examinees with diverse personality traits have different academic performances (Beaujean et al., 2011; Chamorro-Premuzic & Furnham, 2003, 2008; Coenen et al., 2021; Conard, 2006; De Feyter et al., 2012; Furnham et al., 2003; Furnham & Chamorro-Premuzic, 2004; Nye et al., 2015). Sometimes, these personality traits interact with the item-writing flaws and make the assessment unfair and dependent on the examinees' personality (Frary, 1991).

Standardized recommendations for writing items were first proposed by Haladyna and Downing (1989a, 1989b). They reviewed numerous manuals written by lecturers about their own experiences creating items. By compiling the information, they identified common practices and finally collected 31 recommendations (Haladyna et al., 2002). Fourteen out of these 31 initial recommendations were specifically focused on writing alternatives to the correct answer, which are called *distractors* (Haladyna & Rodriguez, 2013). The importance of writing distractors lies in the complexity of making them plausible. This task often consumes the most time when designing an exam. If examinees easily discard the alternatives to the correct answer, the items will lack discrimination making it impossible to distinguish between individuals with high and low content knowledge. This meticulous writing demands considerable effort, which sometimes leads teachers to prefer alternatives that are easier to generate, such as "None-of-the-above" (NOTA; Frey et al., 2005). However, the use of NOTA is often discouraged, despite limited research on the relevance of writing recommendations and inconclusive empirical evidence (Tarrant et al., 2009). Therefore, finding a strategy that reduces the cost of writing items without compromising their quality is highly relevant.

Use of NOTA

In their initial review, Haladyna and Downing (1989b) noted that items with NOTA are more difficult and less discriminatory. Test scores are also less reliable, and criterion validity is negatively affected. They concluded that they had no advantage in this alternative to consider their recommendation, and reinforce this recommendation years later, in their manual (Haladyna & Rodriguez, 2013). Even so, they commented that the ideal approach was to write distractors that provide information to the item and focus on the question.

Nevertheless, contradictory results are still found, though three general issues could be highlighted. First, some authors point out

Corresponding author: Susana Sanz; Email: ssvelasco@comillas.edu

Cite this article: Sanz, S., García, C., & Olmos, R. (2026). Use of Crossed Random-Effects Models to Assess Multiple-Choice Items: An Experimental Study. *The Spanish Journal of Psychology* **29**, e8, 1–10. https://doi.org/10.1017/SJP.2026.10022

© The Author(s), 2026. Published by Cambridge University Press on behalf of Universidad Complutense de Madrid and Colegio Oficial de la Psicología de Madrid. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/SJP.2026.10022 Published online by Cambridge University Press

that item difficulty increases when NOTA is included. Some of them consider this increment as something positive when the test would be overly easy otherwise (Frary, 1991; Rich & Johanson, 1990), whereas others argue that this increment is artificial and unrelated with the examinees' skills (DiBattista et al., 2014; Downing, 2005). Mathematically, including NOTA increases difficulty and reduces discrimination, unless very few items with NOTA are included (Persson, 2023). However, there is a general agreement that the increase in difficulty is more pronounced when NOTA is the correct answer, as opposed to being a distractor (DiBattista et al., 2014; Martínez et al., 2008; Pachai et al., 2015). Some authors also commented on the uncertainty NOTA creates during the assessment as one of the biggest problems for the examinees (Boland et al., 2010; Dochy et al., 2001; Frary, 1991). When defending their use, the recommendation is to include NOTA only when better distractors cannot be created, considering it a superior alternative to a non-functional distractor, as it enhances discrimination (Pachai et al., 2015; Rich & Johanson, 1990). Yet, Little (2023) reports that, when including NOTA as a one of the distractors, neither performance nor confidence is harmed when students evaluate item-by-item, whereas they indicate, later on, that they were less confident when they found NOTA items, which states a contradiction in the two different moments. She suggests that there are beliefs about NOTA that could lead students to think that their performance would worsen, even when they do not have real evidence about that. However, she do not have information about the impact in the perceived harm when NOTA is the correct answer, so student could determine that NOTA was always an incorrect answer, leading to a better performance and more confidence in their answers.

Lastly, some authors discuss the use of NOTA concerning testing effects, meaning the formative power of assessment. When NOTA is the correct answer, the exam does not aid examinees in knowledge acquisition, as they lack access to the correct answer during the test (Blendermann et al., 2020; Jang et al., 2014; Odegard & Koen, 2007). In this regard, NOTA can devaluate the assessment's purpose as an educational tool, as items can be answered correctly without understanding the syllabus (Boland et al., 2010). Additionally, its inclusion may increase guessing, which reduces the reliability of obtained scores (Persson, 2023).

Personality Traits

Similar to NOTA, research on personality shows contradictory results. Chamorro-Premuzic and Furnham (2003) found a positive relationship between Conscientiousness and examination grades and a negative relationship between these grades with Extraversion and Neuroticism. These authors also reported the same positive relationship between academic performance and Conscientiousness and a negative one with Extraversion (Furnham et al., 2003). Both studies considered the overall marks in the first two academic years. They also found a positive correlation between statistics examination grades and Conscientiousness and a negative correlation with Extraversion (Furnham & Chamorro-Premuzic, 2004). Conard (2006) identified a positive relationship between three different measures of academic performance (college GPA, course performance, and class attendance) and Conscientiousness, but found no relationship with other traits. Again, Chamorro-Premuzic and Furnham (2008) found that academic performance, based on overall marks in the first university year, correlated positively with Openness to Experience and Conscientiousness. Beaujean et al. (2011) examined the relationship between personality and reading and math achievement. They found that Conscientiousness and Openness to Experience influenced both reading and math performance, Agreeableness predicted reading achievement, and Extraversion affected math performance. Moreover, De Feyter et al. (2012) found an effect of four out of the five Big Five traits. Agreeableness, Conscientiousness, and Neuroticism positively predicted academic success (measured as the proportion of obtained to attempted credits), while Extraversion showed a negative relationship. Nye et al. (2015) found an effect on academic success (measured as GPA and a Russian unified exam comprising Math, Language, and Social Sciences) for all the traits except Conscientiousness. In their study, they reported a negative relationship with Extroversion and a positive effect on Agreeableness, Neuroticism, and Openness to experience. Finally, Coenen et al. (2021) found that Conscientiousness had a positive relationship with academic performance and Risk Preference had a negative relationship, while they found no significant results for Emotional Stability. Academic performance was measured by grades in a quantitative methods course.

Aim and Objective

Considering this information, the main interest of this article is to explain the use of NOTA and analyze whether there are differences when it is used as the correct answer versus a distractor, to determine if it negatively affects the quality of multiple-choice items. To accomplish this, three different forms of the same statistics items were designed based on the presence of NOTA, and its function within the item, resulting in three different exams. Thus, we could systematize the use of NOTA in the same items but with different roles and compare it with its absence.

Further, while personality traits and academic performance have been linked in numerous studies, most analyses include only correlational data. More complex models could more accurately capture the effect of personality traits on the probability of answering exam questions correctly, which in turn influences academic performance. Although the evidence from the cited studies shows that the personality traits most closely linked to academic performance are Conscientiousness, Extraversion, and Neuroticism, including all Big-Five personality traits and Impulsivity in more refined analyses could reveal the effects of other traits.

Additionally, the relationship between personality and how examinees handle NOTA options could be relevant. As Frary (1991) pointed out, personality might be related to differences in reactions to this type of item, as only some examinees reported increased uncertainty, whereas Little (2023) defended that this is a student's belief but has no real implication in their performances.

Unraveling the true functioning of NOTA and determining if its use could be acceptable under certain circumstances will help lecturers save time and effort in creating items, allowing them to invest it in others. This will lead to exams that better reflect examinee performance, ultimately enhancing validity and fairness in assessment. In this sense, an experimental study is proposed, as the different conditions can be manipulated and controlled to draw stronger conclusions. Moreover, using more complex models could help capture variance related to items and examinees combined, as well as their potential interaction. Crossed random-effects models could model these complex relationships, allowing for the inclusion of covariates related to examinees (such as personality traits) or items (psychometric properties or writing flaws, like NOTA in this case). Further, we also estimate a Linear Logistic Test Model (LLTM; Fisher, 1973) to estimate the general effect of including NOTA as the correct answer or a distractor.

https://doi.org/10.1017/SJP.2026.10022 Published online by Cambridge University Press

The specific hypotheses evaluated are as follows:

1. The presence of NOTA may decrease the probability of answering correctly. This effect will be more noticeable when NOTA is the correct answer than when it is a distractor.
2. Personality traits will affect the probability of answering correctly (Conscientiousness, Emotional Stability and Openness to Experience in a positive direction; Extroversion and Impulsivity in a negative direction).
3. Examinees will exhibit different behavior when answering NOTA items, and this behavior could be influenced by their personalities.

Method

Participants

The examinees were 449 volunteer psychology students at Universidad Autónoma de Madrid (UAM) who signed a consent indicating the tasks of the study and could leave the experiment anytime if they decide. They were rewarded with extra credits that they could exchange in any course associated with the rewards program. However, they were requested to pass the Statistics I course, which is a mandatory course for first-year undergraduates, to get access to the experiment. This project was approved by the research ethics committee of the university.

The main task had three forms, randomly distributed across participants to guarantee the experimental requirements: Form A was answered by 146 examinees, Form B by 156 examinees, and Form C by 147 examinees. A control task was conducted, and some of the data were eliminated due to the detection of disengagement. The control task was a series of questions where participants had to choose a specific answer; otherwise, they were excluded. After this check, 435 examinees (140 for Form A, 152 for Form B, and 143 for Form C) were retained. For Form A, the group included 22 males and 103 females, with a mean age of 20.66 years (SD = 3.79). For Form B, the group included 26 males and 119 females, with a mean age of 20.49 years (SD = 2.17). For Form C, the group included 14 males and 125 females, with a mean age of 19.95 years (SD = 1.76).

The data were collected in two different proctored sessions with a limited time, in order to unify the administration conditions as much as possible. Therefore, the answers from both sessions were linked through an anonymous ID, and the examinees who do not provide data for one of the sessions were removed from the final database. Finally, 409 examinees (125 for Form A, 145 for Form B, and 139 for Form C) were retained. The mean age of the examinees was 20.36 years old (SD = 2.68, min = 17, max = 59). Sixty-two examinees were males, and 347 were females.

Design of the Study

A concept inventory to capture the differences between the use of NOTA was developed. These instruments have been traditionally made to address misconceptions in some fields, especially in physics and chemistry. Halloun and Hestenes (1985) developed the first inventory concept to help the examinees understand Newtonian physics concepts. They demonstrate that it was especially useful to change examinees' previous beliefs about difficult concepts. For this reason and knowing that statistical misconceptions are common and hard to change, these instruments are suitable for the present study. The starting point to create the task was the Statistic Concepts Inventory (SCI), which was developed to assess Engineering examinees' understanding of statistical concepts (Stone et al., 2003). The inventory includes items referred to as probability, descriptive statistics, inferential statistics, and graphical displays. This inventory was adapted for psychology students, resulting in 30 items with three-alternative tests to measure descriptive and inferential statistics. Ten items had three specific options, which were the same for all examinees, as an anchor test. Then, the other 20 items were balanced to finally get three versions of each one: one with three specific options, one with NOTA as the correct answer and two distractors, and one with NOTA as one of the distractors, besides the correct option and another distractor. With these items, three forms of the test were created. One had three specific options for each of the 30 items, and the other two included the 10 anchor items and 20 with the NOTA option, balancing its use as the correct option or one of the distractors between these two different forms. The three versions of the test were randomly distributed to the participants. Figure 1 shows a graphic about the designed task.

An example of item for the three forms is shown in Table 1, and the complete SCI (in Spanish) can be consulted in the Supplemental Appendix.

In terms of personality traits instruments, two different tests were used: a Spanish version of the big-five personality test, OPERAS (Costa & McCrae, 1992; Vigil-Colet et al., 2013), and a Spanish version of Barrat's Impulsiveness Scale (BIS-11; Oquendo et al., 2001; Patton et al., 1995). OPERAS contains 40 personality items, 7 of each big-five personality scale and the rest considering Social Desirability, and respondents must answer in a Likert scale from 1 to 5. Regarding the BIS-11, the scale contains 30 items in which respondents must answer a 4-point frequency scale. Some personal and academic information was also collected. In particular, the participants were asked about their age, gender, and their marks in the Spanish University-access exam and each statistics subject in the degree (Research methods, Statistics I and Statistics II).

Procedure and Data Analyses

The data were collected in two proctored sessions, one for the SCI and, a week later, the personality and personal data.

To start with the analyses, only the 10 items for the anchor test were considered to determine if the three groups that have taken the

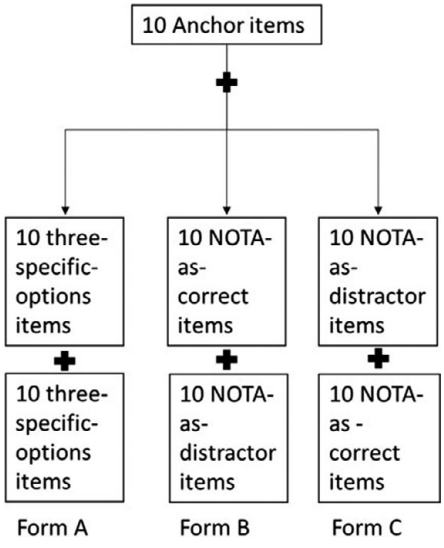


Figure 1. Structure of the three forms of SCI tasks.

https://doi.org/10.1017/SJP.2026.10022 Published online by Cambridge University Press

Table 1. Example of an item in their three versions

The following values correspond to the Intelligence Quotients (IQ) of 7 university students: 104, 103, 108, 111, 108, 106, and 103. By how much the IQ of the most intelligent student could be increased without changing the median?		
Form A	Form B	Form C
(a) It could be increased by any amount.	**(a) It could be increased by any amount.**	(a) It could only be increased by 2 points.
(b) It could only be increased by 2 points.	(b) Any increase would raise the median.	(b) Any increase would raise the median.
(c) Any increase would raise the median.	(c) None of the above.	**(c) None of the above.**

The correct answer is in bold. Please note that Forms B and C are counterbalanced. In 10 items, NOTA is the correct answer in Form B and another 10 times is a distractor, and the same happen in Form C.

three different test versions were comparable in terms of test reliability and items difficulty. To assess measurement invariance across the three groups, a Multiple Indicators, Multiple Causes (MIMIC) model was employed (Jöreskog & Goldberger, 1975). Whereas a multi-group confirmatory factor analysis could account for a more complete invariance study, they can be more sensitive to small group sizes, potentially leading to reduced statistical power to detect non-invariance or estimation difficulties. Alternatively, MIMIC models incorporate group variables directly as predictors of the latent factor and/or the observed item indicators, being more appropriate considering the sample size. Then, the reliability of the complete test in every form and the personality scales were analyzed.

A long format database was used to perform the rest of the analyses, and all the variables were standardized, in order to better compare the coefficients. To model the probability of getting the correct answer for each item, a crossed random-effects analysis was performed. We took a Bayesian approach, as it could be more suitable for complex structures, like crossed random-effects models could have, as it can enhance stability for the estimations. Moreover, we set weak priors, specifically normal distributions with means of zero and standard deviations of three for the mean and standard deviation on items and examinees. Setting weak prior distributions allows us to regularize the estimations without giving information that could bias them, and normal distributions are common when talking about personality traits and items properties.

In these models, items and examinees are the random effects and the different covariates are fixed effects (Janssen et al., 2004). These models offer significant advantages in analyzing complex data structures compared to more simple models. For instance, ANOVAs average one of the factors, or simple regressions consider only one of them, leading to a partial understanding of the explained variability, and sometimes failing to capture all of it (Baayen et al., 2008; Hoffman, 2015; Judd et al., 2012). However, crossed random-effects models incorporate different factors as sources of variability (in this case, items and examinees), overcoming these problems. This leads to manage complex interactions between variables from distinct levels and estimate variances across these levels. They efficiently manage incomplete data, allowing for the use of partial information without discarding complete observations. Additionally, they provide flexibility in including predictors at various levels, enhancing the ability to capture the intricate relationships between variables, which leads to a more detailed understanding of the sources of variability and the dynamics within the data (Judd et al, 2017).

The different items were considered repeated measures from the examinees' knowledge level and personality traits as variables that do not change over time. The dependent variable is the achievement of the correct answer (a dichotomic variable, where 0 is a failure and 1 is an achievement). The differences in difficulty between items and the possible differences between examinees when they deal with NOTA as an option were also considered in the crossed random-effects model.

In particular, the examinees-level effects considered were impulsivity (IM), conscientiousness (CO), agreeableness (AG), emotional stability (ES), openness to experience (OE), and extraversion (EX). As information about the grades of the three courses ("Research Methods," "Statistics I," and "Statistics II") was relatively related, we finally decided to only include information about the course grade that was passed by the larger percentage of students, which was "Research Methods."

Further, the item-level effects were the discrimination of the different options (correct answer, most chosen distractor, and less chosen distractor), NOTA as the correct answer, and NOTA as a distractor.

To further test the effect of NOTA and complement the crossed random-effects model performed, a Bayesian-LLTM analysis was also performed, which have been used in some studies evaluating the effect of the response options (Hagenmüller, 2020; Hohensinn & Baghaei, 2017). LLTM is an extension of the Rasch model (Rasch, 1980), which decomposes the item parameter (in this case, difficulty) into different categories that could increase or decrease the value. Applying this model could help to determine the impact of NOTA as the correct answer and a distractor in general terms, besides the specific question.

The analyses were conducted with Mplus (version 8; Muthén & Muthén, 2017) and R software (version 4.1.2; Development Core Team, 2021), in particular with "brms" (Bürkner, 2021) and "rstan" (Stan Development Team, 2021) packages.

Results

Groups Equivalence

Classical Test Theory

The first step was checking that the three groups were equivalent, so only the 10 anchor items were considered, and the data for the 435 examinees that complete the SCI were used. The score reliability in the three anchor tests was compared (Feldt et al., 1987), and Table 2 shows Cronbach's α for the three forms and their confidence intervals.

Table 2. Cronbach's α for the 10-item forms and confidence intervals (95%)

Form	Cronbach's α [95% CI]
A	.387 [.224, .528]
B	.509 [.384, .618]
C	.421 [.269, .553]

https://doi.org/10.1017/SJP.2026.10022 Published online by Cambridge University Press

Table 3. Difficulty in anchor items in the three forms

	Item 1	Item 5	Item 8	Item 13	Item 16	Item 19	Item 24	Item 26	Item 29	Item 33
Form A	.743	.357	.193	.393	.450	.671	.107	.650	.686	.243
Form B	.724	.303	.204	.414	.434	.632	.145	.664	.684	.401
Form C	.797	.315	.238	.427	.385	.769	.140	.573	.678	.259

The alpha comparison did not show significant differences between the three groups' reliability ($\chi^2_{2;\ .05} = 1.602$, $p = .449$).

Then, the differences in difficulty in the anchor test were checked. Table 3 shows the values in difficulty between the three forms.

A between-subjects ANOVA with Bonferroni multiple-comparison correction was performed for every item, and no significant differences were found, as the corresponding α for a 99% confidence level is 0.003 (F_1: $_{2,\ 432} = 1.141$, $p = .32$; F_5: $_{2,\ 432} = 0.535$, $p = .586$; F_8: $_{2,\ 432} = 0.465$, $p = .628$; F_{13}: $_{2,\ 432} = 0.169$, $p = .844$; F_{16}: $_{2,\ 432} = 0.677$, $p = .509$; F_{19}: $_{2,\ 432} = 3.452$, $p = .033$; F_{24}: $_{2,\ 432} = 0.523$, $p = .593$; F_{26}: $_{2,\ 432} = 1.489$, $p = .227$; F_{29}: $_{2,\ 432} = 0.01$, $p = .99$; F_{33}: $_{2,\ 432} = 5.432$, $p = .005$).

MIMIC Model

Finally, a MIMIC model to assure the groups' equivalence was performed (Jöreskog & Goldberger, 1975; Muthén, 1989), as the sample size could be controversial in terms of analysis stability, and it is not recommended to perform a multiple-groups confirmatory analysis, as MIMIC models entail the analysis of a single covariance matrix. This implies a more parsimonious model favoring the estimation of the invariance model that can otherwise be difficult to estimate (i.e., multiple-groups CFA entails three variance–covariance matrices estimated with categorical data; Brown, 2015). In MIMIC models, the latent factor and the items were regressed onto the form (i.e., the group was taken as a covariate). To assure that performing a MIMIC model is correct, we first fit a one-factor CFA model ($\chi^2_{35;\ 0.05} = 35.769$, $p = .432$; RMSEA = .007; CFI = .995; TLI = .993). Given the results, which indicate a good fit of the data, we proceed with the MIMIC model, which also provides a good fit for the data ($\chi^2_{44;\ 0.05} = 45.279$, $p = .418$; RMSEA = .008; CFI = .991; TLI = .989). The form covariate did not change the factor structure or produce strain areas in the solution (all modification indices were < 10.0). This result is consistent with the previous ones and shows that the indicators are invariant for the three forms.

Complete Forms Analyses

Classical Test Theory

As with the anchor test, the reliability of the complete SCI forms and the personality scales were checked. Table 4 shows the Cronbach's α and the confidence intervals of each measure.

As it can be seen, Cronbach's α values did not show significant differences across the three complete forms but were lower than expected. In terms of personality, all the scales have acceptable reliability, but it is not equal across the scales.

Difficulty index for all the items was also calculated. Table 5 shows the results for the three different forms.

Crossed Random-Effects Model

A first Bayesian crossed random-effects model with all the variables mentioned in the Data Analyses section was conducted. Then, the non-significant variables were eliminated to optimize the model and to explain as much variance as possible with the significant variables.

Table 4. Cronbach's α for the completed forms and personality scales (with confidence intervals)

	Cronbach's α [95% CI]
SCI Form	
A	.605 [.498, .698]
B	.566 [.458, .662]
C	.613 [.514, .700]
Personality scale	
IM	.772 [.739, .803]
ES	.841 [.816, .863]
OE	.776 [.742, .807]
EX	.874 [.854, .892]
CO	.736 [.695, .773]
AG	.688 [.640, .732]

Table 6 shows the results of the optimized model, including all the statistically significant variables.

Related to the items, the significant variables were the item-rest correlation for the most chosen distractor (β = −1.49; 95% CI [−2.06, −0.93]) and the inclusion of NOTA as a correct answer (β = −.24; 95% CI [−0.40, −0.08]) and as a distractor (β = −.44; 95% CI [−0.58, −0.29]). The three variables point to the increase in the probability of failure of the item when they are presented. Statistically significant differences in three of the personality traits were found: extraversion, emotional stability, and conscientiousness. According to the results, extroverted (β = −.06; 95% CI [−0.11, −0.01]) and conscientious (β = −.08; 95% CI [−0.13, −0.02]) people tend to fail more, whereas emotionally stable (β = .07; 95% CI [0.02, 0.13]) people get the correct answer. However, these effects are small. Lastly, the research methods' grades (β = .17; 95% CI [0.13, 0.20]) are useful to predict the probability of getting the correct answer, even when the syllabus of these subjects is not directly related to the content of the SCI.

According to the model, examinees were expected to react to NOTA items in different ways depending on personal variables. Table 7 shows random effects estimation, which in this sense reflects the estimations for the standard deviations associated with NOTA as the correct answer or a distractor. Significant results support the hypothesis that there are differences between examinees when they have to deal with the presence of NOTA as an alternative, as the variance is different among examinees.

Significant results were found for all the values, which means that there are differences between examinees, both when NOTA is the correct answer and a distractor.

https://doi.org/10.1017/SJP.2026.10022 Published online by Cambridge University Press

Table 5. Difficulty for all the items in the three forms

Item	Form A	Form B	Form C
1	.743	.724	.797
2	.636	.592	.664
3	.057	.145	.077
4	.650	.414	.664
5	.357	.303	.315
6	.814	.658	.650
7	.429	.414	.301
8	.193	.204	.238
9	.829	.868	.685
10		Control item	
11	.443	.434	.343
12	.321	.164	.573
13	.393	.414	.427
14	.321	.401	.322
15	.521	.474	.587
16	.450	.434	.385
17	.700	.730	.636
18	.243	.243	.224
19	.671	.632	.769
20	.621	.493	.615
21	.143	.191	.168
22		Control item	
23	.571	.553	.587
24	.107	.145	.140
25	.643	.691	.566
26	.650	.664	.573
27	.621	.467	.490
28	.300	.349	.308
29	.686	.684	.678
30		Control item	
31	.864	.776	.748
32	.929	.770	.755
33	.243	.401	.259

White background: 3-specific-options items; light gray background: NOTA-as-distractor items; dark gray: NOTA-as correct items.

LLTM

A Bayesian LLTM (Fisher, 1973) was performed to check if NOTA as an option resulted significantly, reinforcing the previous conclusions. In this case, to distinguish between the complete item and the item without alternatives, *question* is used.

First, to check the fit of the data, the Rasch model was estimated, and a Bootstrap Goodness-of-Fit analysis using Pearson chi-squared was performed, showing a good fit of the data ($p = .9$, with a 95% of confidence level). This result suggests that the data do not significantly deviate from the model's expectations. Furthermore, the correlation between the difficulty parameters for the Rasch model and the easiness parameter for the LLTM is −0.99. This high correlation confirms that the assumptions of the model have been met.

Table 6. Fixed effects model estimates for correct answer probability (final model, standardized)

	β	Estimation error	95% CI_{Min}	95% CI_{Max}
Intercept	−1.01	0.23	−1.45	−0.58
Item-rest correlation (most chosen distractor)	−1.49	0.29	−2.06	−0.93
NOTA correct	−.24	0.08	−0.40	−0.08
NOTA distractor	−.44	0.07	−0.58	−0.29
Research methods' grade	.17	0.02	0.13	0.20
Extraversion	−.06	0.03	−0.11	−0.01
Emotional Stability	.07	0.03	0.02	0.13
Conscientiousness	−.08	0.03	−0.13	−0.02

Note: CI = confidence interval.

Table 7. Random effects estimates for the final model

	β	Estimation error	95% CI_{Min}	95% CI_{Max}
SD (NOTA correct)	.63	0.08	0.48	0.78
SD (NOTA distractor)	.42	0.09	0.24	0.58
Correlation (NOTA correct – NOTA distractor)	.50	0.20	0.12	0.90

Note: CI = confidence interval; SD = standard deviation.

Table 8 shows the LLTM parameters for every stem, NOTA as the correct answer and NOTA as a distractor. In this case, lower values represent easiness, and higher values represent difficulty.

As it can be seen, the results obtained performing LLTM are similar to the crossed random-effects results, showing that NOTA both as the correct answer ($\eta = 0.20$; 95% CI [0.04, 0.35]) and as a distractor ($\eta = 0.31$; 95% CI [0.16, 0.46]) increase items' difficulty.

Discussion

The main goal of this study was to assess whether crossed random-effects models can capture how examinee and item characteristics influence the probability of correct answers, and ultimately their scores. In particular, we were interested in the effect of flaws in the stems or the alternative answers, as NOTA. These flaws could lead to the evaluation of other variables such as their ability to detect these flaws and guess the correct answer, which would affect the reliability and validity of the results, as we were confounding the explained variance by measuring variables different from the knowledge level in the course.

As the results in the reliability of the scores and items' difficulty indicate that the three groups were equivalent, and this conclusion is supported by the MIMIC model results, we considered that the results related to NOTA were going to be reliable as well. However, higher values of Cronbach's α would be desirable to reinforce the security of the conclusions. The exam difficulty, in combination

https://doi.org/10.1017/SJP.2026.10022 Published online by Cambridge University Press

Table 8. LLTM parameters

η parameter	Estimate	SE	95% CI_{Min}	95% CI_{Max}
Question 1	−1.19	0.12	−1.43	−0.96
Question 2	−0.78	0.12	−1.02	−0.53
Question 3	2.17	0.18	1.84	2.51
Question 4	−0.49	0.12	−0.72	−0.26
Question 5	0.78	0.11	0.57	1.00
Question 6	−1.11	0.13	−1.35	−0.86
Question 7	0.30	0.12	0.08	0.54
Question 8	1.40	0.12	1.15	1.64
Question 9	−1.63	0.14	−1.91	−1.36
Question 10	0.22	0.12	−0.02	0.44
Question 11	0.47	0.12	0.25	0.71
Question 12	0.33	0.11	0.12	0.53
Question 13	0.47	0.12	0.23	0.70
Question 14	−0.29	0.12	−0.53	−0.05
Question 15	0.33	0.10	0.13	0.53
Question 16	−1.04	0.13	−1.29	−0.80
Question 17	1.09	0.13	0.83	1.35
Question 18	−0.83	0.12	−1.07	−0.61
Question 19	−0.47	0.12	−0.70	−0.25
Question 20	1.53	0.14	1.25	1.81
Question 21	−0.49	0.12	−0.71	−0.26
Question 22	1.98	0.15	1.70	2.27
Question 23	−0.76	0.12	−0.99	−0.53
Question 24	−0.59	0.11	−0.80	−0.39
Question 25	−0.32	0.12	−0.56	−0.08
Question 26	0.66	0.12	0.41	0.90
Question 27	−0.80	0.11	−1.02	−0.58
Question 28	−1.61	0.14	−1.88	−1.34
Question 29	−1.74	0.13	−2.01	−1.47
Question 30	0.84	0.12	0.61	1.07
NOTA correct	0.20	0.08	0.04	0.35
NOTA distractor	0.31	0.07	0.16	0.46

Note: CI: confidence interval; SE = standard error.

with the fact that it was a low-stakes assessment, could have affected the value of this indicator, introducing a greater percentage of guessing (Ulitzsch et al., 2020). In this sense, a pilot study could have been beneficial to detect this issue, which is one limitation of the study. Moreover, the purpose of a concept inventory could be different from the purpose of a conventional exam, which could lead to differences in terms type of knowledge being assessed (more specific in an exam and wider in a concept inventory). In this sense, maybe the structure of a conventional exam could have been more suitable, which would be considered in future studies.

About the item-level results, we found that the item-rest correlation of the most chosen distractor was the most impactful variable, which seems logical considering that values of this indicator are expected to be high and negative. The higher, more positive value in this indicator, the greater the probability of failing the item. NOTA as an option also seems to harm the examinees' performance, being greater when NOTA is one of the distractors, and it seems to interact with examinees. These results were consistent when performing a Bayesian LLTM, but contradict the first hypothesis, as it was expected that the probability of getting the correct answer would be lower when NOTA is the correct answer and not a distractor, as seen in other studies (DiBattista et al., 2014; Martínez et al., 2008; Pachai et al., 2015). This could be explained by the fact that in the other studies, the overall difficulty is lower than in this one. More research studies are needed, but the conservative recommendation would be to avoid including NOTA options, as they reduce the probability of getting the correct answer. Even if we consider the increment in difficulty as something desirable, this increment can be considered artificial, as we are introducing more uncertainty in the items, and the positive effects could be related with the deletion of a non-plausible distractor, rather than the inclusion of NOTA (Butler, 2018; Pachai et al., 2015). Further, the inclusion of NOTA does not provide any positive evaluation effects, as the examinees do not have access to the correct answer when taking the exam and NOTA is the correct answer, and prevent knowledge acquisition (Blendermann et al., 2020; Jang et al., 2014; Odegard & Koen, 2007). Moreover, the difficulty increment is not equal across the examinees, so the assessment turns out to be less fair when NOTA is used. We hypothesized that the inclusion of NOTA could lead to an increase in difficulty for the most competent individuals, as they could detect some unintended details in the response alternatives and doubt about the correct answer, while less competent individuals could disregard them. However, we would like to explore in-depth the examinees' variables that affect the different behavior in the presence of NOTA.

Further, about the examinee-level effects, there are small but significant effects on Extraversion, Emotional Stability, and Conscientiousness. A specific measurement of the effect size cannot be given, as researchers have not already provided an accurate way to report it in crossed random-effects models (Correll et al., 2022). Greater values in Extraversion and Conscientiousness lead to a greater probability of failing the item, in contrast to Emotional Stability, which leads to a greater probability of getting the correct answer. Extraversion and Emotional Stability results coincide with the results found in other studies (Chamorro-Premuzic & Furnham, 2003; De Feyter et al., 2012; Furnham et al., 2003; Furnham & Chamorro-Premuzic, 2004; Nye et al., 2015). The Conscientiousness results were surprising, as the opposite effect was expected, due to the results in other papers (Beaujean et al., 2011; Chamorro-Premuzic & Furnham, 2003, 2008; Coenen et al., 2021; Conard, 2006; De Feyter et al., 2012; Furnham et al., 2003; Furnham & Chamorro-Premuzic, 2004). However, in the case of these studies, academic performance is measured by taking into account the marks in many subjects, which could give a more general perspective, as these marks have greater consequences on examinees' academic trajectory. The different perspectives could have also affected the generalization of the results. However, our experimental focus and the aim to link the information about personality traits and the use of NOTA did not allow us to consider the other authors' focus. Furthermore, Nye et al. (2015) considered in their study that the education system and the cultural context can explain some differences between studies. Furnham and Monsen (2009) found no relationship between math ability and conscientiousness. SCI includes statistical concepts items that the examinees

https://doi.org/10.1017/SJP.2026.10022 Published online by Cambridge University Press

have not specifically prepared for the task, so it required math ability to answer and could explain the differences in the results. Finally, examinees that have obtained a high grade in the Research methods subject also seem to have a greater probability of getting the correct answer, even when the SCI items were unrelated to this subject's syllabus.

This work is not exempt from limitations. As mentioned, Cronbach's α values are lower than desirable, so better results on reliability could lead to more stable conclusions. These results are also influenced by the low-stakes assessment conditions, as the examinees earned extra credit for participation, but there is no impact in their real evaluations, which would be more realistic but is unethical, as the expectation was that NOTA harmed academic performance. Furthermore, the great difficulty of the exam could also lead to a greater failure percentage than expected in this kind of assessment. Easier items should be incorporated into the exam to reduce the general difficulty and increase the variability, so this could lead to avoid range restriction in the examinees' scores and, as a result, increase the reliability. On the other hand, the replacement of the distractors with NOTA has depended on the original SCI data (Stone, 2006), and the logic was using the most useful distractors, replacing the ones that were worse psychometric properties. We thought that more "natural alternatives" are, indeed, easier to generate and were the first in which the lectures think when writing an exam, so eliminating these alternatives sounded more implausible. However, replacing different distractors could have led to different results in terms of difficulty and discrimination, even if we are using the same items in the three versions of the test.

Some other future lines are exploring the Conscientiousness results in future works to find out if there are cultural differences in the higher education trajectories in Spain, compared to other countries, or if the type of contents in SCI could lead to differences in performance (Furnham & Monsen, 2009; Nye et al., 2015). In this sense, exploring the different Conscientiousness facets could provide some answers, as the OPERAS does not include them, and other works have provided evidence of different correlations in these facets with academic performance (Chamorro-Premuzic & Furnham, 2003).

Furthermore, exploring what happen when we consider the omissions would be important, as they could also explain the performance of the examinees and provide more information about the examinees' behavior when they must deal with items (Ulitzsch et al., 2020). Future studies should work on IRT models that include the effect of the omissions on the probability of getting the correct answer, the personality traits, and the items' characteristics. These models would also include the discrimination differences between items, and a guessing parameter, which could provide more information about the item's functioning.

Finally, although differences between examinees have been found when NOTA is used as a response, it is still necessary to detect the reason it is not related to the specific personality traits included in this work. In this sense, as the anchor items did not vary between forms, statistical power to detect interactions between NOTA and personality traits could be reduced. Moreover, the small sample size could also lead to fail to detect the specific personality traits related to the effect of NOTA.

Some other traits, such as the intolerance of uncertainty, could be relevant in this context and help us explain the differences (Bryan et al., 2021).

In conclusion, the results point out that including NOTA reduces the fairness in the assessment, introducing variability that does not depend on the examinees' knowledge. Personality traits are also important and trying to reduce their impact in the evaluation, for instance, providing a quiet space and enough time for the evaluation, or making more stimulant activities for extroverted examinees to retain the important concepts could reduce the harmful impact in the final assessment. Furthermore, crossed random-effect models provide higher-quality results, as they can consider a greater amount of variability related to items and examinees.

Supplementary Material. The supplementary material for this article can be found at http://doi.org/10.1017/SJP.2026.10022.

Data Availability Statement. Data are available upon request due to privacy and ethical considerations.

Author Contribution. Conceptualization: all authors; Data curation: S.S.; Formal analysis: all authors; Investigation: S.S., C.G.; Methodology: all authors; Project administration: C.G.; Resources: C.G., R.O.; Software: S.S., R.O.; Supervision: C.G., R.O.; Validation: C.G.; Writing—original draft: S.S.; Writing—review and editing: all authors.

Competing Interests. None declared.

References

Baayen, R. H., **Davidson, D. J.**, & **Batesm, D. M.** (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, **59**(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Beaujean, A. A., **Firmin, M. W.**, **Attai, S.**, **Johnson, C. B.**, **Firmin, R. L.**, & **Mena, K. E.** (2011). Using personality and cognitive ability to predict academic achievement in a young adult sample. *Personality and Individual Differences*, **51**(6), 709–714. https://doi.org/10.1016/j.paid.2011.06.023

Blendermann, M. F., **Little, J. L.**, & **Gray, K. M.** (2020). How "none of the above" (NOTA) affects the accessibility of tested and related information in multiple-choice questions. *Memory*, **28**(4), 473–480. https://doi.org/10.1080/09658211.2020.1733614

Boland, R. J., **Lester, N. A.**, & **Williams, E.** (2010). Writing multiple-choice questions. *Academic Psychiatry*, **34**(4), 310–316. https://doi.org/10.1176/appi.ap.34.4.310

Bryan, E. M., **Beadel, J. R.**, **McLeish, A. C.**, & **Teachman, B. A.** (2021). Assessment of intolerance of uncertainty: Validation of a modified anagram task. *Journal of Behavior Therapy and Experimental Psychiatry*, **73**, 101671. https://doi.org/10.1016/j.jbtep.2021.101671

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Publications.

Bürkner, P. C. (2021). Bayesian item response modelling in R with brms and Stan. *Journal of Statistical Software*, **100**(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition*, **7**(3), 323–331. https://doi.org/10.1016/j.jarmac.2018.07.002

Chamorro-Premuzic, T., & **Furnham, A.** (2003). Personality traits and academic examination performance. *European Journal of Personality*, **17**, 237–250. https://doi.org/10.1002/per.473

Chamorro-Premuzic, T., & **Furnham, A.** (2008). Personality, intelligence and approaches to learning as predictors of academic performance. *Personality and Individual Differences*, **44**(7), 1596–1603. https://doi.org/10.1016/j.paid.2008.01.003

Coenen, J., **Golsteyn, B. H. H.**, **Stolp, T.**, & **Tempelaar, D.** (2021). Personality traits and academic performance: Correcting self-assessed traits with vignettes. *PLoS One*, **16**(3), e0248629. https://doi.org/10.1371/journal.pone.0248629

Conard, M. A. (2006). Aptitude is not enough: How personality and behavior predict academic performance. *Journal of Research in Personality*, **40**, 339–346. https://doi.org/10.1016/j.jrp.2004.10.003

Correll, J., **Mellinger, C.**, & **Pedersen, E. J.** (2022). Flexible approaches for estimating partial eta squared in mixed-effects models with crossed random

factors. *Behavior Research Methods*, **54**, 1626–1642. https://doi.org/10.3758/s13428-021-01687-2

Costa, P., & **McCrae, R.** (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.

De Feyter, T., **Caers, R.**, **Vigna, C.**, & **Berings, D.** (2012). Unravelling the impact of the Big Five personality traits on academic performance: The moderating and mediating effects of self-efficacy and academic motivation. *Learning and Individual Differences*, **22**, 439–448. https://doi.org/10.1016/j.lindif.2012.03.013

DiBattista, D., **Sinnige-Egger, J.**, & **Fortuna, G.** (2014). The "none of the above" option in multiple-choice testing: An experimental study. *The Journal of Experimental Education*, **82**(2), 168–183. https://doi.org/10.1080/00220973.2013.795127

Dochy, F., **Moerkerke, G.**, **De Corte, E. D.**, & **Segers, M.** (2001). The assessment of quantitative problem-solving skills with "none of the above"-items (NOTA items). *European Journal of Psychology of Education*, **16**(2), 163–177. https://doi.org/10.1007/BF03173023

Downing, S. M. (2005). The effects of violating standard item writing principles on test and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, **10**, 133–143. https://doi.org/10.1007/s10459-004-4019-5

Feldt, L. S., **Woodruff, D. J.**, & **Salih, F. A.** (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, **11**, 93–103. https://doi.org/10.1177/014662168701100107

Fisher, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, **37**, 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Frary, R.B. (1991) The None-of-the-Above Option: An Empirical Study. *Applied Measurement in Education*, **4**(2), 115–124, http://dx.doi.org/10.1207/s15324818ame0402_2

Frey, B. B., **Petersen, S.**, **Edwards, L. M.**, **Pedrotti, J. T.**, & **Peyton, V.** (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, **21**(4), 357–364. https://doi.org/10.1016/j.tate.2005.01.008

Furnham, A., & **Chamorro-Premuzic, T.** (2004). Personality and intelligence as predictors of statistics examination grades. *Personality and Individual Differences*, **37**(5), 943–955. https://doi.org/10.1016/j.paid.2003.10.016

Furnham, A., **Chamorro-Premuzic, T.**, & **McDougall, F.** (2003). Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and Individual Differences*, **14**(1), 49–66. https://doi.org/10.1016/j.lindif.2003.08.002

Furnham, A., & **Monsen, J.** (2009). Personality traits and intelligence predict academic school grades. *Learning and Individual Differences*, **19**(1), 28–33. https://doi.org/10.1016/j.lindif.2008.02.001

Hagenmüller, B. (2020). On the impact of the response options' position on item difficulty in multiple-choice-items. *European Journal of Psychological Assessment*, **37**(4), 290–299. https://doi.org/10.1027/1015-5759/a000615

Haladyna, T. M., & **Downing, S. M.** (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, **2**(1), 37–50. https://doi.org/10.1207/s15324818ame0201_3

Haladyna, T. M., & **Downing, S. M.** (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, **2**(1), 51–78. https://doi.org/10.1207/s15324818ame0201_4

Haladyna, T. M., **Downing, S. M.**, & **Rodriguez, M. C.** (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, **15**(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5

Haladyna, T. M., & **Rodriguez, M. C.** (2013). *Developing and validating test items*. Routledge.

Haladyna, T. M., **Rodriguez, M. C.**, & **Stevens, C.** (2019). Are multiple-choice items too fat? *Applied Measurement in Education*, **32**(4), 350–364. https://doi.org/10.1080/08957347.2019.1660348

Halloun, I., & **Hestenes, D.** (1985). The initial knowledge state of college physics students. *American Journal of Physics*, **53**(11), 1043–1055. https://doi.org/10.1119/1.14030

Hoffman, L. (2015). *Longitudinal Analysis. Modeling withinperson fluctuation and change*. Routledge.

Hohensinn, C., & **Baghaei, P.** (2017). Does the position of response options in multiple-choice tests matter? *Psicológica*, **38**(1), 93–109.

Jang, Y., **Pashler, H.**, & **Huber, D. E.** (2014). Manipulation of choice familiarity in multiple-choice testing support a retrieval practice account of the testing effect. *Journal of Educational Psychology*, **106**(2), 435–447. https://doi.org/10.1037/a0035715

Janssen, R., **Schepers, J.**, & **Peres, D.** (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 188–212). Springer.

Jöreskog, K. G., & **Goldberger, A. S.** (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, **70**(351a), 631–639. https://doi.org/10.2307/2285946

Judd, C. M., **Westfall, J.**, & **Kenny, D. A.** (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, **103**(1), 54–69. https://doi.org/10.1037/a0028347

Judd, C. M., **Westfall, J.**, & **Kenny, D. A.** (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, **68**, 601–625. https://doi.org/10.1146/annurevpsych-122414-033702

Little, J. L. (2023). Does using none-of-the-above (NOTA) hurt students' confidence? *Journal of Intelligence*, **11**(8), 157. https://doi.org/10.3390/jintelligence11080157

Martínez, R. J., **Moreno, R.**, **Martín, I.**, & **Trigo, M. E.** (2008). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psicothema*, **21**(2), 326–330.

Muthén, L.-K., & **Muthén, B. O.** (2017). *MPLUS (version 7).* [Computer Software]. https://www.statmodel.com/

Muthén, B. O. (1989). Latent variable modelling in heterogeneous populations. *Psychometrika*, **54**(4), 557–585. https://doi.org/10.1007/BF02296397

Nye, J. V. C., **Orel, E.**, & **Kochergina, E.** (2015). *Big five personality traits and academic performance in Russian universities.* Research Paper No. #WP BRP 10/PSY/2013. https://doi.org/10.2139/ssrn.2265395

Odegard, T.N. & **Koen, J.D.** (2007) "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory*, **15**(8), 873–885. https://doi.org/10.1080/09658210701746621

Oquendo, M. A., **Baca-García, E.**, **Graver, R.**, **Morales, M.**, **Montalvan, V.**, & **Mann, J. J.** (2001). Spanish adaptation of the Barratt Impulsiveness Scale (BIS-11). *The European Journal of Psychiatry*, **15**(3), 147–155. https://doi.org/10.1016/j.ijchp.2015.07.002

Pachai, M. V., **DiBattista, D.**, & **Kim, J. A.** (2015). A systematic assessment of "None of the above" on a multiple-choice test in a first-year psychology classroom. *The Canadian Journal for the Scholarship of Teaching and Learning*, **6**(3), 2–14. http://doi.org/10.5206/cjsotl-rcacea.2015.3.2

Patton, J. H., **Stanford, M. S.**, & **Barrat, E. S.** (1995). Factor structure of the Barrat Impulsiveness Scale. *Journal of Clinical Psychology*, **51**(6), 768–774. https://doi.org/10.1002/1097-4679(199511)51:6%3C768::AID-JCLP2270510607%3E3.0.CO;2-1

Persson, R. A. X. (2023). *Multiple-choice test items with "none of the above" as an alternative.* PsyArXiv. https://doi.org/10.31234/osf.io/w7cnr

R Core Team. (2021). *R: A language and environment for statistical computing (version 4.1.2).* [Computer Software]. R Foundation for Statistical Computing.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests (Expanded edition, with foreword and afterword by B. D. Wright)*. University of Chicago Press.

Rich, C. E., & **Johanson, G. A.** (1990). An Item-Level Analysis of "None of the Above". *Paper presented at the annual meeting of the American Education Research Association*, Boston, MA.

Stan Development Team (2021). *RStan: The R interface to Stan. R package version 2.21.3*. Author.

Stone, A., **Allen, K.**, **Rhoads, T. R.**, **Murphy, T. J.**, **Shehab, R. L.**, & **Saha, C.** (2003). The statistic concept inventory: A pilot study. *Proceedings of the 33rd Annual Frontiers in Education. FIE 2003*, Westminster, CO, USA, 2003, pp. T3D-1. https://doi.org/10.1109/FIE.2003.1263336

Stone, A. (2006). *A psychometric analysis of the Statistic Concept Inventory.* [Doctoral dissertation]. University of Oklahoma. http://hdl.handle.net/11244/1013

Tarrant, M., **Ware, J.**, & **Mohammed, A. M.** (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, **9**, 40. https://doi.org/10.1186/1472-6920-9-40

Ulitzsch, E., **von Davier, M.**, & **Pohl, S.** (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, **73**(S1), 83–112. https://doi.org/10.1111/bmsp.12188

Vigil-Colet, A., **Morales-Vives, F.**, **Camps, E.**, **Tous, J.**, & **Lorenzo-Seva, U.** (2013). Development and validation of Overall Personality Assessment Scale (OPERAS). *Psicothema*, **25**(1), 100–106. https://10.7334/psicothema2011.411

https://doi.org/10.1017/SJP.2026.10022 Published online by Cambridge University Press