frontiers | Frontiers in Education

TYPE Brief Research Report
PUBLISHED 17 October 2024
DOI 10.3389/feduc.2024.1442047

OPEN ACCESS

EDITED BY
Jinnie Shin,
University of Florida, United States

REVIEWED BY
Maura Pilotti,
Prince Mohammad bin Fahd University, Saudi Arabia
Morgan Les DeBusk-Lane,
Gallup, United States

*CORRESPONDENCE
Séverin Lions
✉ severin.lions@ciae.uchile.cl

RECEIVED 01 June 2024
ACCEPTED 01 October 2024
PUBLISHED 17 October 2024

CITATION
Soto C, Lions S, Ortega G, Arjona M, Blanco MP and Dartnell P (2024) The arrangement of response options in multiple-choice test items: verticality is not always better.
Front. Educ. 9:1442047.
doi: 10.3389/feduc.2024.1442047

The arrangement of response options in multiple-choice test items: verticality is not always better

Consuelo Soto[1], Séverin Lions[1,2]*, Gabriel Ortega[1], Melissa Arjona[1], María Paz Blanco[1] and Pablo Dartnell[1,3,4]

[1]Center for Advanced Research in Education (FB0003), Institute of Education, Universidad de Chile, Santiago, Chile, [2]Departamento de Evaluación, Medición y Registro Educacional (DEMRE), Universidad de Chile, Santiago, Chile, [3]Center for Mathematical Modeling (FB210005), Universidad de Chile, Santiago, Chile, [4]Department of Mathematical Engineering, Universidad de Chile, Santiago, Chile

Multiple-choice tests are widely used to measure learning outcomes. Consequently, constructing high-quality test items is critical, and many authors have advanced item-writing guidelines. One frequently mentioned guideline is to arrange the response options vertically. However, evidence to support this recommendation is scarce and has only been obtained for items with text-based options. This study aimed at understanding whether the arrangement of options affects performance at solving items with large-sized options, such as graphs and pictures, using objective and subjective measures. Fifty-seven high-school students completed a multiple-choice science and mathematics test with 24 four-choice items, options being graphs or pictures presented in one of four arrangements: vertical without page break, vertical with page break, Z, and inverted N. Response accuracy, response time, and perceived difficulty were obtained for each item. Subsequently, students participated in a cognitive interview about their experiences, practices, perceptions, and beliefs regarding the arrangement of options. Objective measures show that the arrangement of options hardly affected performance, the only effect being that vertical condition with page break resulted in significantly longer response times. Subjective measures show that most students favored the vertical arrangement they consider more common but negatively perceived vertical condition with page break and considered squared arrangements (Z, inverted N) to facilitate visual exploration and comparison between options, as opposed to the vertical arrangement. Results suggest that the vertical arrangement does not offer clear advantages over squared arrangements for items with large-sized options.

KEYWORDS

multiple-choice, item-writing guidelines, item formatting, response options, spatial arrangement

1 Introduction

Multiple-choice tests are among the most effective educational assessment tools and are widely used to measure students' learning outcomes (Gierl et al., 2017; Moreno et al., 2015). In many countries, such as the United States, Japan, China, South Korea or Chile, performance on this type of test becomes critical, as it determines students' likelihood of being admitted to the careers and universities of their choice (Durán del Fierro, 2019; Moreau, 2015). Given their essential role, multiple-choice tests should accurately measure learning outcomes (Downing, 2005). Consequently, item-writing flaws must be avoided (Tarrant and Ware, 2012). In

particular, items should be formatted to optimize legibility and promote text exploration so that formatting issues do not hinder item solving (Haladyna and Rodriguez, 2013).

Many guidebooks have provided recommendations to help item writers draft high-quality test items (Lions et al., 2024). Most guidebooks include formatting guidelines, such as "place response options in a consistent order" (Lions et al., 2022). One format characteristic frequently mentioned in these guidebooks is the arrangement of response options, item writers being generally invited to present options in a vertically-displayed list (Moreno et al., 2006). Based on data from the two most cited reviews on item-writing guidelines (Haladyna and Downing, 1989; Haladyna et al., 2002), most textbooks identified as dealing with options arrangement (19/29, that is, 66%) recommend using vertical format. Consistently, the vertical arrangement has been broadly adopted in the testing industry and is now universally used.

Several authors have ventured hypotheses regarding why a vertical display of response options might be optimal. Verticality is thought to maximize text segmentation and thus facilitate comparisons between options (Moreno et al., 2004, 2006, 2015). It is also thought to favor a more efficient visual scanning of each option (Considine et al., 2005; Reynolds et al., 2006). According to studies on the perceptual span, i.e., the portion of effective visual information extracted per eye fixation (Frey and Bosse, 2018; Rayner, 1998; Rayner et al., 2010), vertical arrangement of options might promote an efficient scanning of each option separately. However, although the verticality guideline is empirically testable, evidence supporting it is scarce (Haladyna and Downing, 1989; Haladyna et al., 2002). Just three studies are found regarding this issue: two empirical (Bendulo et al., 2017; Follman et al., 1969) and one on examinees' perceptions (Oyzon et al., 2016). In Follman et al. (1969), 80 college students were randomly assigned to one option arrangement condition (vertical, horizontal) and asked to respond to a 53-item comprehension test. Although test performance was globally higher when options were vertically displayed, the option arrangement effect was not significant. Bendulo et al. (2017), based on McConkie and Rayner's (1975) study on the perceptual span, experimentally presented more option arrangement conditions (vertical, horizontal, Z, inverted N), but found no significant effect of options arrangement again, this time on scores of a 60-item general culture test responded by 176 students. These two studies do not offer any clear-cut empirical support for verticality. Finally, Oyzon et al. (2016) administered a survey on options arrangement to 261 university students and found that students had a strong preference for the vertical arrangement they perceived as a facilitator of options text exploration.

The optimal arrangement of options might depend on their content. When options are pictures, graphs, or diagrams, as is frequent in science or mathematics, standardized test developers sometimes choose to display these large-sized options in a squared configuration (in Z or inverted N), instead of a vertical one (see for example items from Scholastic Aptitude Test, International Benchmark Tests, or PISA). This may suggest that the vertical arrangement might not be optimal for all cases. On occasions, squared arrangements allow large-sized options to be kept on the same page, which has been suggested to be important (Wood et al., 2006). Scanning back and forth while reading an item is time-consuming (Chenevey, 1988), and avoiding page breaks might enable easy reading and analysis of item content (Taylor et al., 1978). Since previous research on options arrangement has been conducted on items with text-only options, the question arises as to whether the arrangement of response options affects the resolution of multiple-choice items when options are pictures or graphs (and thus large), and whether a vertical arrangement may indeed be the optimal way to organize options in this case.

In this study, a classroom experiment was conducted to evaluate whether the arrangement of large-sized options (pictures or graphs) affected students' performance when solving multiple-choice items. After completing the experimental task, students were individually interviewed to gather information on their experiences, practices, perceptions, and beliefs about options arrangement's influence (regarding the particular task of this study and in general). The study collected objective and subjective measures to better understand whether item-solving and test outcomes are affected by option arrangement and why. Research questions were: Does any option arrangement improve test outcomes as compared to the other ones when options are pictures or graphs? If this is the case, which particular ones and why? Do students better value squared arrangements in this context? Should item-writing guides not recommend the vertical arrangement when options are pictures and/or graphs? Should these guides warn against page breaks?

Based on the literature mentioned above, the following hypotheses arise for the case of large-sized options: (a) arrangement of options will have little effect (if any) on test performance (Bendulo et al., 2017; Follman et al., 1969), (b) the vertical arrangement will be reported to make comparing options easier (Moreno et al., 2004, 2006, 2015), (c) the vertical arrangement will be perceived as a facilitator of the visual scanning of options (Considine et al., 2005; Frey and Bosse, 2018; Oyzon et al., 2016; Reynolds et al., 2006; Rayner, 1998; Rayner et al., 2010), (d) students will report the vertical arrangement as their favorite (Oyzon et al., 2016), (e) page-break will negatively affect students' performance and perception (Chenevey, 1988; Taylor et al., 1978; Wood et al., 2006).

2 Method

2.1 Design and task

The experimental task consisted of answering a 24-item four-choice science and mathematics test in which the arrangement of items' response options was carefully manipulated. All participants solved the same items, with each participant solving these items in a different random order. Participants solved six items of each one of the four experimental conditions: vertical without page break, vertical with page break, Z, and inverted N (see Figure 1). Conditions were counterbalanced across administered test forms to generate a fully crossed design. Response accuracy, response time, and perceived difficulty were registered and subsequently analyzed. Participants underwent an individual cognitive interview after completing the test. A descriptive analysis of their experiences, practices, perceptions, and beliefs regarding the arrangement of options was performed. Data collection was conducted on two different days so all participants could take the test and be interviewed on the same day. All research protocols were approved by the Ethics Committee of the Faculty of Social Sciences of the *Universidad de Chile*.

10.3389/feduc.2024.1442047

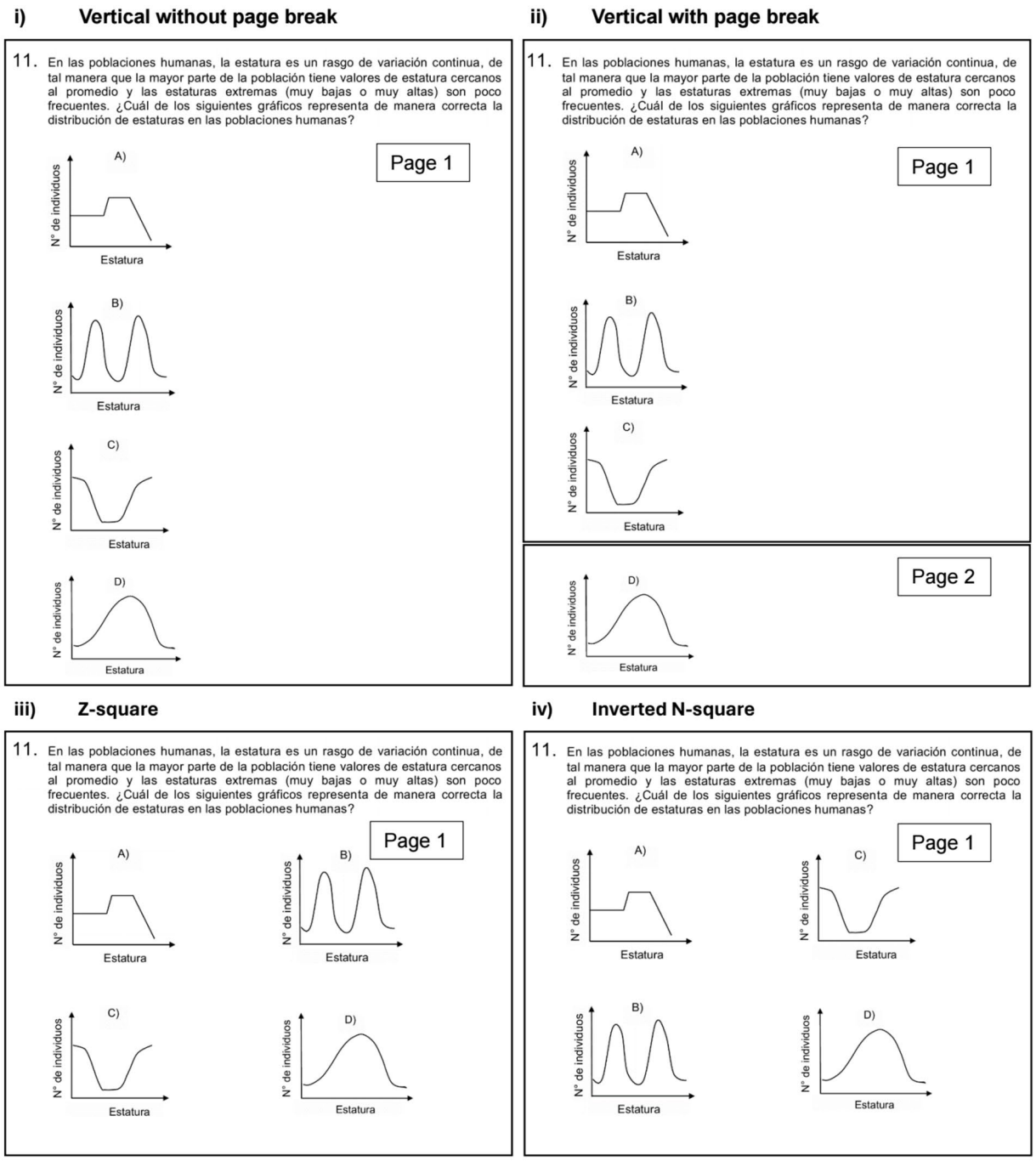

FIGURE 1
The four experimental conditions. (i) Vertical without page break, (ii) vertical with page break, (iii) Z, and (iv) Inverted N.

2.2 Sample

Fifty-seven Chilean high school students (68% female, mean age = 16.45 ± 0.50 years) participated in this study. They were all native Spanish speakers with normal or corrected-to-normal vision and no neurodevelopmental disorders. All participants and their legal representatives signed a written informed consent authorizing their data to be used for research purposes. Most participants (81.8%) had no specific training in solving multiple-choice tests. The remaining participants (18.2%) reported receiving extracurricular academic support to prepare for the Chilean national university admission tests (PAES), which probably included high exposure to the multiple-choice format.

2.3 Materials

2.3.1 Stimuli

The stimuli used in the classroom experiment were items provided by DEMRE (Departamento de Evaluación, Medición y Registro

Educacional), the state institution responsible for developing and administering national Chilean university admission tests. All were from the official 2016–2020 DEMRE's item bank and had thus previously been validated by DEMRE through expert reviews and field testing. DEMRE also provided stimuli's item characteristics so that stimuli could be selected based on the following criteria: (a) response options were pictures or graphs, and (b) item difficulty was not too high (facility index <0.2) or too low (facility index >0.8). Since DEMRE's items had five options, the distractor with the lowest response rate was removed to obtain four-option items (because previous studies on options arrangement have used four-option items and it is the most commonly used number of options). Participants reported almost no prior exposure to the stimuli (only 8 participants reported being possibly familiar with just one item).

Test forms included 24 items (six were mathematics, twelve were physics, four were chemistry, and two were biology). They were presented in physical format as single-sided letter-size pages (width = 21.59 cm, height = 27.94 cm). An initial page containing test instructions was included. At the top and the bottom of each page, a space was included to record time (hour, minute, and seconds) before the beginning to solve the item (starting time) and after answering the item (completion time), respectively. Additionally, a scale of perceived difficulty was included for participants to complete after solving each item, with four levels ranging from 1 = Very easy to 4 = Very difficult.

2.3.2 Cognitive interview script

Semi-structured cognitive interviews were conducted based on a 3-section script. The first section aimed at determining whether participants noticed that different option arrangements could be observed in the test they just took and then asked participants to report their personal experiences, practices, and perceptions regarding vertical and squared patterns of options (arrangement noticing, scanning order, personal preference, prior exposure, perceived best, page break noticing, and page break perception). The second section revealed the four experimental conditions to participants and probed into participants' beliefs regarding favorite arrangements among students, ease of options comparison, ease of text exploration, and most common arrangements used in multiple-choice tests. The last section inquired about possible clinical diagnoses that might have affected the reading task (dyslexia, attention deficit hyperactivity disorder, visual impairment), previous intensive training in solving multiple-choice tests, and familiarity with the test items used in this study (see the cognitive interview script in Supplementary material), that might have required to exclude participants or trials from the analyses.

2.4 Procedure

Data were collected over two days using the same procedure. The multiple-choice test was administered at the beginning of the school day (8.05 am.–9.35 am.) in a classroom setting. Instructions were projected on the blackboard for participants to read before taking their tests. Participants were instructed to answer items as accurately and rapidly as possible and to meticulously register both starting/completion times and perceived difficulty. Official time was permanently projected on the blackboard during the whole session so that participants could use it for time reports. The test lasted about 40 min on average. Participants were given a snack upon completion, and individual one-on-one interviews were conducted from 9.50 am.–1.00 pm. Two trained interviewers conducted interviews in parallel in two different quiet rooms. The interview lasted eight minutes on average. Each interviewer covered every section of the cognitive interview script, and the students' raw answers were recorded during the interview on an Excel sheet. One voluntary closing session was conducted to present the study results to participants one month after collecting the data to thank them for participating.

2.5 Data processing and analyses

All participants answered all test items (totaling 1,368 responses). Response times were computed based on each item's completion and starting times. Missing values for response times (11 missing data) and perceived difficulty reports (17 missing data) were replaced using an iterative imputation method based on Random Forest, specifically the MissForest algorithm (Stekhoven and Bühlmann, 2012). Negative response times were treated as missing data.

The effect of option arrangement on students' performance was studied by conducting a mixed-effects logistic regression to examine the influence of the experimental condition on response accuracy and conducting two separate mixed-effects linear regressions to examine the influence of the experimental condition on response time and perceived difficulty. For these regressions, the experimental condition was treated as a fixed-effect factor with an intercept, vertical without page break was taken as reference condition, and subjects and items were incorporated into the models as random effects to account for the variability of subjects and items (see formulas in Supplementary method). An additional mixed-effects logistic regression for response accuracy, with response time included in the model, was conducted to evaluate the existence of a possible speed-accuracy trade-off. The fulfillment of statistical assumptions was checked for each model before running it. In the response time model, data was log-transformed, and outliers were imputed using the MissForest algorithm so that data fulfilled all normality assumptions (see Supplementary method). Additionally, data from the interviews were used to build participant subgroups *a posteriori* (e.g., participants who reported reading squared-displayed options in Z versus inverted N order), and the performance outcomes of the participant subgroups were contrasted and analyzed.

Fifty-five students out of the 57 who took the multiple-choice test were individually interviewed. Raw verbal reports were encoded during the interviews by each interviewer. Subsequently, two raters independently encoded participants' responses based on previously and consensually defined response categories (mean agreement = 95%, range = 71–100%). A third research team member validated these response categories before raters performed this coding process. A descriptive analysis of response frequencies was implemented for each one of the addressed dimensions (section 1: arrangement noticing, scanning order, personal preference, prior exposure, perceived best, page break noticing, and page break perception; section 2: favorite arrangements among students, ease of comparison between options, ease of text exploration, most common arrangements used in multiple-choice tests). The mentioned advantages and disadvantages of each option arrangement and reasons given to support each preference were also co-encoded and analyzed in frequency.

3 Results

3.1 Performance

Mean response accuracy, mean response time, and mean perceived difficulty for participants were 35.3% (±10.0%), 85.5 s (±24.1 s), and 2.9 (±0.4), respectively. Response accuracy was close to the one expected by random selection, but the time participants spent on tasks and the high perceived difficulty suggested that this low performance was not due to careless responding. Instead, participants probably did their best at answering items that were highly challenging for them. Further evidence of this was that all items were correctly answered by at least some participants (the items' percentage of correct responses ranged from 7.0 to 91.2%) and that the more difficult the items were perceived, the lower was the response accuracy and the longer was the response time (mean response accuracy and mean response time were 72.3, 48.2, 31.9, and 24.5%; and 65.1 s, 77.5 s, 89.4 s, and 97.5 s, respectively, for perceived difficulty of level 1, 2, 3 and 4, see more details in Supplementary results).

No differences in response accuracy or perceived difficulty were observed between the experimental conditions (all $p_s >= 0.19$ associated with fixed-effect coefficients). However, response times were longer in vertical with page break than in vertical without page break condition (intercept = 0.176, SE = 0.051; vertical with page break: $\beta = 0.082$, SE = 0.036, 95% CI [0.011, 0.152], $t(1292.03) = 2.270$, $p = 0.023$, $\eta p^2 = 0.010$, $d_{Cohen} = -0.203$, see Table 1 and Supplementary results for a complete report of statistical indicators), suggesting that page break was associated with time loss. No speed-accuracy trade-off was observed, showing that this time loss was not associated with accuracy gains. All these results indicated that the only effect of option arrangement on performance detected in the experiment was that of page breaks. No further significant differences were observed when analyzing performance outcomes of specific participant subgroups (see Supplementary results).

3.2 Experiences, practices, and perceptions

Most participants (87.3%) noticed the presence of several option arrangements. Most of them (90.9%) reported usually exploring vertically-arranged options sequentially from A to D (participants who reported other practices followed different, miscellaneous reading patterns). When faced with squared arrangements, more participants reported exploring options in Z order (60%) than in inverted N order (25.5%), context-dependent order (10.9%), or X order (3.6%). Most participants (81.8%) stated having a favorite arrangement: preference for vertical arrangement was more frequent (47.3%) than preference for squared arrangements (27.3%); a few participants reported an item-dependent preference (7.3%). Almost all participants (98.2%) reported unbalanced prior exposure to arrangements. Reporting exposure to the vertical arrangement as prevailing (67.3%) was more frequent than reporting exposure to squared arrangements as prevailing (5.5%). Several participants (25.5%) declared that exposure depended on the nature of options (text or images) because squared arrangements were more frequently used in math items (7.3%) or items with graphs as options (7.3%). Many participants (67.3%) maintained that some arrangements do seem better suited than others. However, vertical arrangement and squared arrangements were equally perceived as the best arrangement (29.1% for both). Additionally, some participants (9.1%) reported that it all depends on the nature of options or subject matter. Finally, all participants noticed that for some items, options were not on the same page; most participants (81.8%) mentioned that page breaks have negative consequences (e.g., they come as a surprise or are even annoying or confusing; see Figure 2).

TABLE 1 Performance results.

	Experimental condition			
Performance measure	Vertical without page break	Vertical with page break	Z	Inverted N
Response accuracy	34.5 (47.6)	37.7 (48.5)	33.6 (47.3)	35.4 (47.9)
Response time	83.7 (56.2)	93.6 (89.9)	83.3 (65.8)	81.6 (51.2)
Perceived difficulty	2.9 (0.8)	2.9 (0.8)	2.9 (0.8)	2.8 (0.9)

Values are presented as mean (standard deviation).

3.3 Beliefs

Most participants (96.4%) considered that students favor one particular option arrangement. Vertical arrangement was identified as students' favorite by more participants (52.7%) than squared arrangements (36.4%, i.e., 18.2% for Z, 7.3% for inverted N, and 10.9% for any of them). Also, most participants (96.3%) agreed that some arrangements make it easier to compare options. Squared arrangements (Z, inverted N, or both) were considered to favor comparison more frequently (83.4%, i.e., 16.7% for Z, 7.4% for inverted N, and 59.3% for any of them) than vertical arrangement (9.3%). Similarly, most participants (85.2%) agreed that some arrangements promote exploring text more efficiently, and squared arrangements were reported to improve exploration of options by more participants (50.0%, i.e., 18.5% for Z, 1.9% for inverted N, and 29.6% for any of them) than vertical arrangement (31.4%). Finally, all participants agreed that some arrangements are more common than others. Most participants (80%) considered that the most common arrangement is vertical (16.4% said this depends on the nature of options, and only two participants mentioned squared arrangement as the most common).

Participants adduced many reasons for their preferences, some being mentioned by just one or two participants or perceived as advantages equally applying to vertical and squared arrangements (see Figure 3). Nevertheless, some features emerged as a unique or more frequent benefit of one specific arrangement. The vertical arrangement of options was predominantly mentioned as beneficial inasmuch as it is the most common and provides lateral space to write comments or notes next to each option. As for squared arrangements, they were predominantly reported as making it easier to get a gist of all options and compare them. No disadvantages unique (or more frequent) to one particular arrangement emerged.

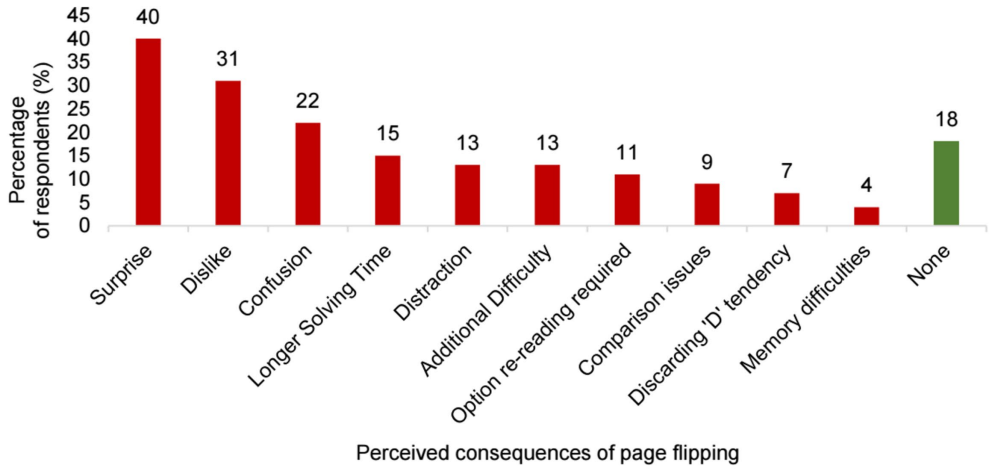

FIGURE 2
Perceived consequences of page flipping. Data is presented as the percentage of participants mentioning each consequence.

4 Discussion

This study analyzed how the arrangement of response options impacts students' performance at answering multiple-choice items with large-sized options, such as graphs or pictures, through objective and subjective measures. Results suggest that students read the vertically-displayed options in sequential order and the squared-displayed options in Z rather than in inverted N direction, that students more frequently favor vertical arrangement (which they deem to be more common and to provide more space for writing annotations), and that students consider that squared arrangements favor option exploration and comparing options. Despite these differential experiences, practices, perceptions, and beliefs, options arrangement was found to hardly affect item solving and performance (the only observed effect was that page break made resolution slower, even if this effect is minimal).

The lack of option arrangement effects on response accuracy could be explained by the fact that the task difficulty was very high and may have masked any existing effect on the probability of correct response (Rice et al., 2012). However, since no effect was observable on any objective performance measures, despite variables not considered in previous studies (response time and perceived difficulty) being analyzed and a within-subject design being used (Charness et al., 2012), a simpler explanation is that this study's results on response accuracy extend those from previous empirical studies (Follman et al., 1969; Bendulo et al., 2017) and support the conclusion that the arrangement of options does not significantly affect performance, independently of options content.

The fact that participants were high-school students and that many preferred the vertical arrangement supports and extends previous findings reporting that vertical arrangement is more popular than horizontal arrangement among university students (Oyzon et al., 2016). However, some cognitive interview results challenged previous conceptions about the vertical arrangement: Contrary to possible predictions based on previous studies (Considine et al., 2005; Frey and Bosse, 2018; Moreno et al., 2004, 2006, 2015; Rayner, 1998; Rayner et al., 2010; Reynolds et al., 2006), in this study squared arrangements, not vertical ones, were most frequently reported to favor both visual exploration of options and making comparisons between options. These deviations from expectations could explain why the vertical arrangement's hypothesized (but not demonstrated) benefits were not observable in this study and might indicate that these predicted benefits do not apply when options are not text.

This study's results confirm that page breaks may hinder the resolution of multiple-choice items (Chenevey, 1988; Taylor et al., 1978). Page breaks made item solving slower and were consistently perceived by students as undesirable for generating confusion, or at the very least surprise, and making comparing options more difficult. Just a few existing item-writing guides recommend maintaining the whole item content on the same page (e.g., Wood et al., 2006). Future guides might add this recommendation, which turned out to be relevant, to their guideline list.

Several pathways for future research can be outlined based on present results. Adding time constraints to the task instead of allowing self-pacing, or administering computer-based instead of paper-based tests might allow more accurately capturing the time spent on the task (Rosenman et al., 2011). Employing eye-tracking techniques might also be helpful (Bendulo et al., 2017), making it possible to compare scan paths linked to different option arrangements or students with different experiences, practices, perceptions, or beliefs on option arrangements. Used along with data from interviews or surveys, it might help identify the mechanisms underlying possible performance effects (ease of comparison between options, ease of option exploration, preference, or high prior exposure). More generally, studying how item-format concerns affect students with learning difficulties seems to be highly important (Kettler et al., 2009; Roelofs, 2019).

This study examines the adequacy of one of the most frequently suggested formatting recommendations in item-writing guides (Haladyna and Downing, 1989; Haladyna et al., 2002). Results show that examinees consider formatting factors to affect their item-solving processes but cast doubts on the verticality guideline. Constructing high-quality multiple-choice items is crucial to obtain valid, reliable, and fair measures of learning. Thus, the other item-writing guidelines with insufficient empirical support should be experimentally challenged so that item-writing becomes an evidence-based activity as soon as possible.

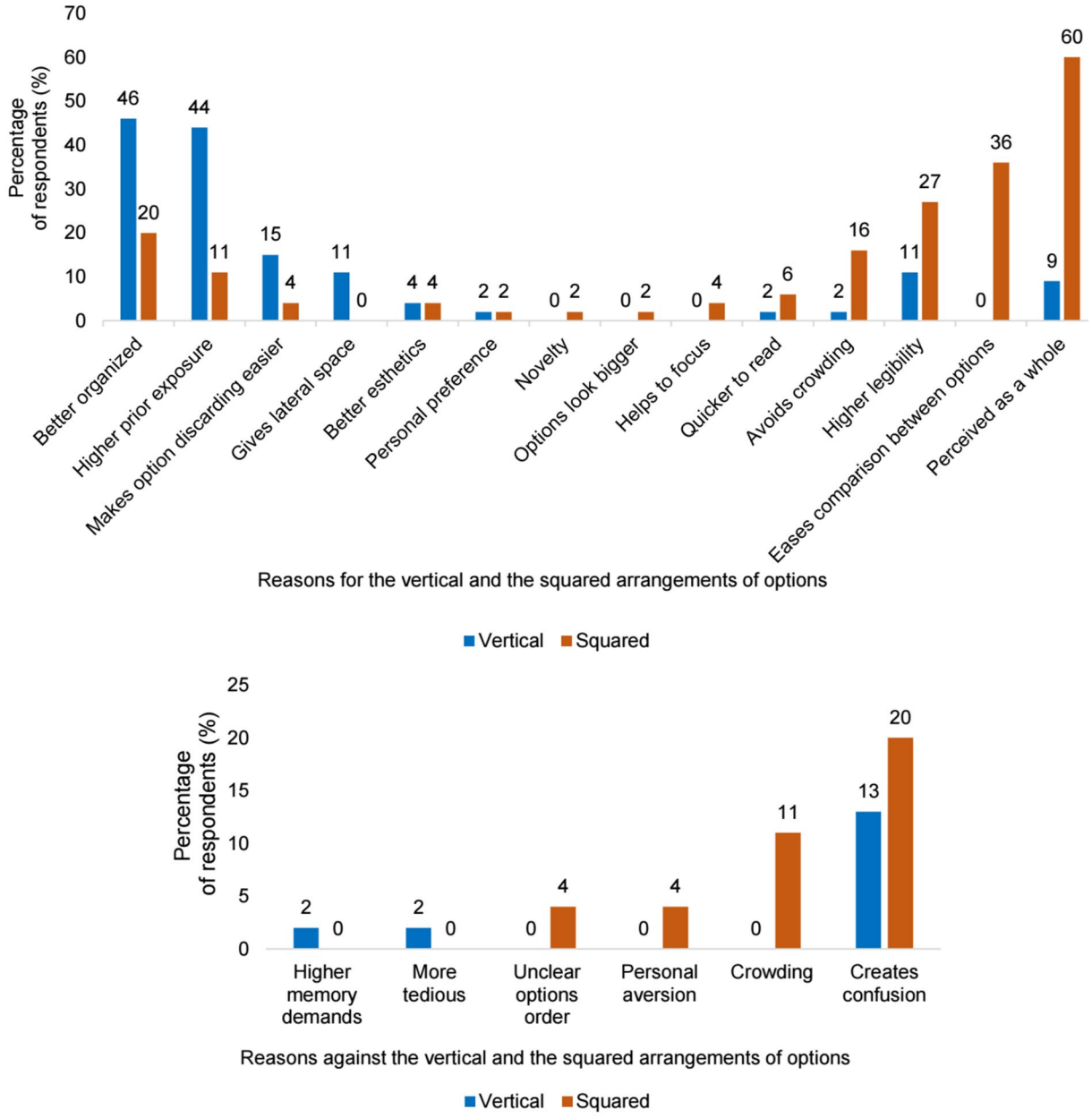

FIGURE 3
Reasons for and against the vertical and the squared arrangements of options.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Ethics Committee of the Faculty of Social Sciences of the Universidad de Chile. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

CS: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. SL: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. GO: Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. MA: Investigation, Writing – original draft, Writing – review & editing.

MB: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. PD: Funding acquisition, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by grants ANID/PIA/Basal Funds for Centers of Excellence FB0003 (Center for Advanced Research in Education) and FB210005 (Center for Mathematical Modeling).

Acknowledgments

We thank Francisca Manríquez Maulén, headmistress of Colegio San José de la Montaña, for her collaboration with this study and DEMRE for providing the math and science items used as stimuli. We also thank Camilo Quezada Gaponov for editing this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2024.1442047/full#supplementary-material

References

Bendulo, H. O., Tibus, E. D., Bande, R. A., Oyzon, V. Q., Macalinao, M. L., and Milla, N. E. (2017). Format of options in a multiple choice test Vis-a-Vis test performance. *Int. J. Eval. Res. Educ.* 6, 157–163. doi: 10.11591/ijere.v6i2.7594

Charness, G., Gneezy, U., and Kuhn, M. A. (2012). Experimental methods: between-subject and within-subject design. *J. Econ. Behav. Organ.* 81, 1–8. doi: 10.1016/j.jebo.2011.08.009

Chenevey, B. (1988). Constructing multiple-choice examinations: item writing. *J. Contin. Educ. Nurs.* 19, 201–204. doi: 10.3928/0022-0124-19880901-05

Considine, J., Botti, M., and Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian* 12, 19–24. doi: 10.1016/S1322-7696(08)60478-3

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Adv. Health Sci. Educ.: Theory Pract.* 10, 133–143. doi: 10.1007/s10459-004-4019-5

Durán del Fierro, F. (2019). Pruebas estandarizadas para el acceso a la educación superior en Chile: performatividad y subjetividad de los estudiantes. *Calidad en la Educación* 50, 180–215. doi: 10.31619/caledu.n50.723

Follman, J., Lowe, A. J., and Miller, W. (1969). "Typeface and multiple choice option format" in Reading: Process and pedagogy–nineteenth yearbook of the National Reading Conference. eds. G. B. Schick and M. M. May, vol. 19, 135–140.

Frey, A., and Bosse, M. L. (2018). Perceptual span, visual span, and visual attention span: three potential ways to quantify limits on visual processing during reading. *Vis. Cogn.* 26, 412–429. doi: 10.1080/13506285.2018.1472163

Gierl, M. J., Bulut, O., Guo, Q., and Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review. *Rev. Educ. Res.* 87, 1082–1116. doi: 10.3102/0034654317726529

Haladyna, T. M., and Downing, S. M. (1989). A taxonomy of multiple choice item-writing rules. *Appl. Meas. Educ.* 2, 37–50. doi: 10.1207/s15324818ame0201_3

Haladyna, T. M., and Rodriguez, M. C. (2013). Developing and validating multiple-choice test items. *4th* Edn. London: Routledge.

Haladyna, T. M., Downing, S. M., and Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Appl. Meas. Educ.* 15, 309–333. doi: 10.1207/S15324818AME1503_5

Kettler, R. J., Elliott, S. N., and Beddow, P. A. (2009). Modifying achievement test items: a theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody J. Educ.* 84, 529–551. doi: 10.1080/01619560903240996

Lions, S., Blanco, M. P., Dartnell, P., Monsalve, C., Ortega, G., and Lemarié, J. (2024). Item-writing guidelines on response options placement: A systematic review. *Applied Measurement in Education*.

Lions, S., Monsalve, C., Dartnell, P., Blanco, M. P., Ortega, G., and Lemarié, J. (2022). Does the response options placement provide clues to the correct answers in multiple-choice tests? A systematic review. *Applied Measurement in Education*, 35,133–152. doi: 10.1080/08957347.2022.2067539

McConkie, G. W., and Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Percept. Psychophys.* 17, 578–586. doi: 10.3758/BF03203972

Moreau, J. (2015). 1915-2015: cent ans de QCM [1915-2015: one-hundred years of MCQs]. *Bulletin de l'Association des Professeurs de Mathématiques de l'Enseignement Public* 516, 516–525.

Moreno, R. M., Martínez, R. J., and Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema* 16, 490–497.

Moreno, R., Martínez, R. J., and Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodol. Eur. J. Res. Methods Behav. Soc. Sci.* 2, 65–72. doi: 10.1027/1614-2241.2.2.65

Moreno, R., Martínez, R. J., and Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema* 27, 388–394. doi: 10.7334/psicothema2015.110

Oyzon, V., Bendulo, H., Tibus, E., Abalajen-Bande, R., and Macalinao, M. (2016). Preference of students on the format of options in a multiple-choice test. *Int. J. Eval. Res. Educ.* 5:292. doi: 10.11591/ijere.v5i4.5956

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422. doi: 10.1037/0033-2909.124.3.372

Rayner, K., Slattery, T. J., and Bélanger, N. N. (2010). Eye movements, the perceptual span, and reading speed. *Psychon. Bull. Rev.* 17, 834–839. doi: 10.3758/PBR.17.6.834

Reynolds, C. R., Livingston, R. B., and Willson, V. (2006). Measurement and assessment in education. London: Allyn & Bacon/Pearson Education.

Rice, S., Geels, K., Hackett, H. R., Trafimow, D., McCarley, J. S., Schwark, J., et al. (2012). The harder the task, the more inconsistent the performance: a PPT analysis on task difficulty. *J. Gen. Psychol.* 139, 1–18. doi: 10.1080/00221309.2011.619223

Roelofs, E. (2019). "A framework for improving the accessibility of assessment tasks" in Theoretical and practical advances in computer-based educational measurement. Methodology of educational measurement and assessment. eds. B. Veldkamp and C. Sluijter (Cham: Springer).

Rosenman, R., Tennekoon, V., and Hill, L. G. (2011). Measuring bias in self-reported data. *Int. J. Behav. Health Res.* 2, 320–332. doi: 10.1504/IJBHR.2011.04

Stekhoven, D. J., and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi: 10.1093/bioinformatics/btr597

Taylor, H., Greer, R. N., and Mussio, J. (1978). Construction and use of classroom tests: A resource book for teachers (ED190609). ERIC: British Columbia Department of Education, Victoria University.

Tarrant, M., and Ware, J. (2012). A framework for improving the quality of multiple-choice assessments. *Nurse Educ.* 37, 98–104. doi: 10.1097/NNE.0b013e31825041d0

Wood, T., Cole, G., and Lee, C. (2006). Developing multiple choice questions for the RCPSC certification examinations. Ottawa: Office of Education, Royal College of Physicians and Surgeons of Canada.