Check for updates

Article

LANGUAGE TESTING

Language Testing
1–21
© The Author(s) 2020

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0265532220917316
journals.sagepub.com/home/ltj
SAGE

The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test

Franz Holzknecht
University of Innsbruck, Austria

Gareth McCray
Keele University, UK

Kathrin Eberharter
Benjamin Kremmel
Matthias Zehentner
University of Innsbruck, Austria

Richard Spiby
Jamie Dunlea
British Council, UK

Abstract

Studies from various disciplines have reported that spatial location of options in relation to processing order impacts the ultimate choice of the option. A large number of studies have found a primacy effect, that is, the tendency to prefer the first option. In this paper we report on evidence that position of the key in four-option multiple-choice (MC) listening test items may affect item difficulty and thereby potentially introduce construct-irrelevant variance.

Two sets of analyses were undertaken. With Study 1 we explored 30 test takers' processing via eye-tracking on listening items from the Aptis Test. An unexpected finding concerned the amount of processing undertaken on different response options on the MC questions, given their order. Based on this, in Study 2 we looked at the direct effect of key position on item difficulty in a sample of 200 live Aptis items and around 6000 test takers per item.

Corresponding author:
Franz Holzknecht, University of Innsbruck, Innrain 52, 5th floor, 6020, Innsbruck, Austria.
Email: franz.holzknecht@uibk.ac.at

The results suggest that the spatial location of the key in MC listening tests affects the amount of processing it receives and the item's difficulty. Given the widespread use of MC tasks in language assessments, these findings seem crucial, particularly for tests that randomize response order. Candidates who by chance have many keys in last position might be significantly disadvantaged.

Keywords
Assessing listening, response processes, eye-tracking, linear mixed effects modelling, multiple-choice, ordering effects, primacy effect

Assessing listening is a complex endeavour in which a multitude of factors can affect task difficulty. These factors can be related to both the listeners themselves as well as the listening assessment task (Brunfaut, 2016). Listener-related factors include linguistic characteristics such as language proficiency (Vandergrift, 2006), lexical knowledge (Andriga et al., 2006), background knowledge (Macaro, Vanderplank, & Grahams, 2005), knowledge and use of listening strategies (Vandergrift & Goh, 2012), working memory capacity (Kormos & Sáfár, 2008; Brunfaut & Révész, 2015), or affective dimensions such as anxiety (MacIntyre & Gardner, 1991; Elkhafaifi, 2005) and motivation (Vandergrift, 2005). Task-related factors, on the other hand, encompass characteristics of the listening text including, but not limited to, linguistic complexity (Révész & Brunfaut, 2013) or speed of delivery (Rosenhouse, Haik, & Kishon-Rabin, 2006), and features of the assessment task such as the number of plays (Field, 2015; Holzknecht, 2019) or note-taking (Carrell, 2007). Another crucial task-related factor that can impact task difficulty is response format.

One of the most common response formats in language assessment, and in educational assessment more generally (Butler, 2018), is that of multiple-choice (MC) items. MC items usually consist of a question (or stem) and a number of response options, out of which test takers need to choose the correct answer. Generally, only one response option is correct (often referred to as the "key"), with the other options serving as distractors. MC items are also popular in assessing listening (Green, 2017), and they tend to be easier in terms of item difficulty compared to open-ended formats (In'nami & Koizumi, 2009). However, research on the particular idiosyncrasies of MC items in listening assessment has been sparse.

With the present study, we attempted to fill this gap by investigating response order effects in four-option MC listening test items. We were interested in whether the position of the key affects the difficulty of the item. In particular, we set out to do the following: (1) explore the effect of the location of response options in MC listening test items and the probability of that response being chosen; (2) attempt to explain the mechanisms behind this effect at a person level, if it is found to exist; and (3) draw conclusions and make recommendations about best practice in the construction of listening items to minimize bias in test score interpretation. Before outlining the study, we review relevant literature on this topic according to the following broad categories: ordering effects in general, ordering effects in MC testing, and ordering effects in MC language testing.

Ordering effects

Researchers from diverse disciplines, such as psychology, animal behaviour, travel research, or marketing, have found that when people (and other animals) are presented with several options from which to choose, the spatial location of the individual options in relation to processing order can have an impact on the final choice. A large number of studies in this strand of research have found evidence for a *primacy effect*, that is, the tendency to prefer the first concept or object encountered. For instance, as discussed by Carney and Banaji (2012, p. 1), primacy influences how well things are remembered (Insko, 1954; Miller & Campbell, 1959; Pineno & Miller, 2005), how attached people and other animals are to others (Bolhuis & Bateson, 1990; Johnson, 1992), how strongly people associate themselves with groups (Greenwald, Pickrell, & Farnham, 2002), how persuasive arguments are (Jersild, 1928; Knower, 1936), or how decisively impressions are influenced (Asch, 1946; Jones, Rock, Shaver, Goethals, & Ward, 1968; Krosnick & Alwin, 1987).

However, studies have also found that positional preference in choosing between individual options depends on the type of judgement involved. Christenfeld (1995) showed that when presented with a number of objectively identical options, people generally prefer the middle options in grocery shopping, in toilet selection (for men), and in choosing between four identical symbols in a row, but the last option when deciding on a route through a maze or when planning a route on a map. Other studies have found that performers in the Eurovision Song Contest and in international figure skating contests are judged more favourably when they appear later (Bruine de Bruin, 2005), that travellers booking hotels online prefer the top and bottom listings (Ert & Fleischer, 2016), or that food choices placed at the top and at the bottom of menus are more popular than choices in the middle (Dayan & Bar-Hillel, 2011).

Carney and Banaji (2012) argued that these different results could be explained by the degree of automaticity of the judgements involved. Based on findings of their own and of previous research, they proposed that decisions involving automatic processing are prone to a primacy effect, but "when controlled processing is possible, other influences can (as they rationally should) override the automatic reliance on the first" (Carney & Banaji, 2012, p. 4).

More related to language testing, Winke and Lim (2015) provided strong evidence for primacy effects in the use of a rating scale for grading students' writing performances. In their eye-racking study, raters displayed a clear left-to-right bias in that they would focus longer on the criteria displayed towards the left. Ballard (2017) partly replicated Winke and Lim's study, and her findings confirm this primacy effect. In addition, Ballard found that the raters would also consider the criteria on the left more important, and if they skipped the reading of a criterion altogether, the skipping was much more likely to happen when the criterion was placed on the right rather than on the left.

Ordering effects in MC testing

Ordering effects have also been investigated in relation to key position in MC testing, as they could potentially introduce construct-irrelevant variance into interpretation of test

scores. If a systematic response order bias were to be found in MC testing, it could mean that individual test takers might be unfairly disadvantaged to a certain degree. For example, if the above-mentioned primacy effect also played a role in MC tests, questions with the key in first position might be answered correctly more often than questions with the key in last position. Consequently, when response options are randomized (as is common practice among many testing boards), test versions with a large number of keys in last position might lead to a higher number of incorrect answers. Studies in this area have been conducted mostly in relation to testing knowledge domains such as psychology, chemistry, or trivia (among others) and have revealed mixed results, as we discuss in the following sections.

Studies that found no effect of response position on item difficulty. Whether option ordering matters has been investigated off and on for a long time. For example, in 1963, Marcus randomly assigned 434 psychology students to one of four groups. Each group (of 104 to 113 students) took one version of a 100-item four-option MC psychology achievement test. The key position was randomly distributed, and it was ensured that the key for each item appeared in different positions across the four test versions. Marcus reported that in terms of the percentage of observed correct responses, no statistically significant bias in relation to key position emerged.

Fast-forward more than 40 years, and similar work in revealed similar results. Taylor (2005) had 60 psychology students take three versions of a 30-item, four-option MC psychology achievement test, within which key distribution was different for each version. In version 1, keys were distributed equally across the four positions. In version 2, 40% of keys were in in the first position, 40% in the second position, and 10% in the third and fourth position. In version 3, this was inverted (10% in the first and second position, and 40% in the third and fourth position). Taylor found no effect on item difficulty in relation to key position and suggested that balancing the key might not be as important as widely believed.

These two studies, however, display a number of notable limitations. The first concerns the sample size. With a total of only 113 candidates per group in Marcus (1963) or 60 in Taylor (2005), potential practically significant effects might not have emerged because the implied powers of the experiments were low. In addition, the studies were primarily cross-sectional, with each group split up for the comparative analyses (one group for each test version). The power to look at variables or factors with potentially small, but possibly important, effects can be improved by increasing the number of observations. This can be done by (a) increasing the sample size, or (b) increasing the number of observations through a within-subjects design (having all learners participate in all conditions). Finally, neither of the studies was conducted within the context of the assessment of L2 learning, which makes it difficult to interpret the findings for language testing purposes.

Studies that found an effect of response position on item difficulty. Other research reported an effect of response order on item difficulty, however, none of these studies were conducted within the context of L2 language testing. Cizek (1994), for example, investigated

the effect of correct response placement in an experimental study. During a medical certification examination with 200 items, the participants (380 in group 1 and 379 in group 2), who were all "graduates of medical specialty residency training programs" (Cizek, 1994, p. 11), were given two different answer forms for a 20-item MC segment. Each item had to be answered based on a projected visual stimulus and respondents could select from 30 possible answers. Although the sequence of items and length of display was kept constant, the sequence of responses on the answer form was scrambled by the researcher to create two forms. Despite relatively small numbers in terms of items and group sizes, four out of 20 items displayed statistically significant differences in difficulty. However, Cizek was not able to detect any predictable patterns in item difficulty differences and concluded that no linear relationship between correct response placement and item difficulty could be established based on this data set.

In another study researching a more conventional MC format, Attali and Bar-Hillel (2003) found that test takers were more likely to choose middle positions when guessing. They looked at the performances of about 4000 candidates taking an Israeli university entrance admissions test measuring various scholastic abilities. Their data consisted of 161 MC items, each taken by at least 220 candidates, divided into two groups. One group received the original test (with the response options in their original positions 1, 2, 3, and 4), whereas the other group received the same test with a different response order for each item (2, 1, 4, and 3 instead of 1, 2, 3, and 4). The authors focused on wrong answers, assuming that these would be guesses, and found that, on average, wrong answers in middle positions were chosen 3% more often than wrong answers in extreme positions. In addition, when the key was placed in either of the two extreme positions, the number of correct responses decreased by 3% and item discrimination (biserial) increased by 0.05 points. This effect was bigger on more difficult items. Analyses of live-test data on more than 4500 items and several thousand candidates confirmed the results. The authors concluded that MC tests with many keys in middle positions are slightly easier and less discriminating than tests with many keys in the two extreme positions. However, the authors did not look at individual response positions, nor did they specify which types of items are mostly affected by middle bias.

Contrary to the findings by Attali and Bar-Hillel, a number of studies reported an effect in line with the primacy effect discussed above; that is, test takers prefer earlier options to later ones. Clark (1956), for example, inferred that in five-option MC tests, about 10% of test takers did not read the last two responses. He analysed several thousand candidates' wrong answers on a scholastic aptitude test, a psychology test, and a mental ability aptitude test, controlling for response order, and found that, throughout these tests, the last two responses were chosen 10% less often as a wrong answer than the first three.

Similarly, Fagley (1987) reported a significant positional bias towards earlier responses in four-option MC tests for 10% of test takers. In her investigation, 60 candidates took a 32-item test on television trivia and a 28-item test on learning skills. Fagley found a statistically significant bias for early responses for six candidates in terms of chosen wrong answers. No significant bias was detected for the sample as a whole. This, however, might again be related to the relatively small sample size of the study.

Tellinghuisen and Sulikowski (2008) also reported a primacy effect in their investigation on the effect of response order in a 38-item four-option MC chemistry test. They analysed test scores of candidates who took the same test twice, once in August (697 students) and once in December (half of the students). On both occasions, two versions of the test were administered, differing in item order and response order. When comparing the results of the two administrations, the authors found that for the 12 items where the difference in correct answers between the two occasions exceeded 6%, 10 were items where the key was in an earlier position (i.e., these 10 items were significantly easier when the key appeared earlier).

Ordering effects in MC language testing

To our knowledge, only two investigations have looked at response order bias in MC tests in the field of language testing. In an unpublished study employing an experimental research design (i.e., response position was experimentally manipulated creating different versions of the test), Sonnleitner, Guill, and Hohensinn (2016) found a significant first position bias for a four-option MC vocabulary test taken by 10-year-olds. For this test and population, the same items (with the same distractors) were answered correctly significantly more often with the key in the first position as compared to the fourth position. However, the authors did not find significant effects for an MC reading comprehension test taken by the same students, nor for reading comprehension tests taken by students aged 14 years or older. In another study, Hohensinn and Baghaei (2017) looked at response order effects in the context of the Iranian National University Entrance Test for English studies: a four-option MC test including items on grammar, vocabulary, and reading comprehension. Hohensinn and Baghaei showed that items were answered correctly slightly less often as the key moved to later positions. However, they concluded that this effect was small and that random distribution of answer options is a valid practice (Hohensinn & Baghaei, 2017, p. 107).

Potential causes of primacy in MC listening tests. We were not able to find studies that have examined primacy effects in the selection of responses to listening test items. This is surprising, as it could be argued that listening tests might be more prone to such an effect than reading, vocabulary, or grammar tests, or tests of knowledge domains. Whereas most of the support for a primacy effect in the literature reviewed above comes down to preference for certain positions when given a range of choices, in listening tests the preference of an earlier option could be further exacerbated by cognitive demands. This is because listening effort can be linked to attentional capacity (Strauss & Francis, 2017). In listening tests, candidates not only have to engage in multi-modal processing – listening to the text while simultaneously reading the questions and answers, as well as matching potential answers to the questions – but also pay attention to the specific characteristics of the text, such as phonology, accents, prosodic features, speech rate, or hesitations (Buck, 2001). In addition, for some listening tests the audio files are played only once, without the chance to hear them again, which puts a further restraint on candidates (Field, 2015; Holzknecht, 2019). For example, in the current versions of widely

used high-stakes listening test such as TOEFL, IELTS, TOEIC, PTE Academic, or GEPT, all listening texts are played only once.

Thus, listening test takers might be more reliant on the order of the responses than reading test takers when deciding on answers to MC items, as they have less processing capacity available to carefully read all of the responses. Such an effect would likely be aggravated for lower language proficiency test takers, as they would take more time to read the options than higher proficiency test takers. As shown by Winke and Lim (2014), less proficient (and more anxious) candidates take significantly longer to process answer options than more proficient (and less anxious) candidates. For these reasons, with the time pressures of many listening comprehension tests, language learners may not be able to read properly the options presented in later positions. It is, therefore, reasonable to assume that in MC listening tests, responses presented higher up on the screen (for computerized tests) or on the test paper (for paper-and-pencil tests) might be more easily accessible to test takers and item difficulty could be influenced more strongly by such an effect than in reading tests or tests of knowledge domains.

The present study

If primacy does play a role in MC language testing, as some of the findings presented above seem to suggest, language test developers would need to rethink their practices to avoid introducing test method-related construct-irrelevant variance into their scores. Given the inconclusive findings on response order bias in relation to MC testing of knowledge domains, the limited number of studies addressing this issue in regard to language testing, as well as the lack of research in relation to listening assessment, it seems prudent to investigate this further. The current paper attempts to shed light on this by looking at response order bias in an MC listening test using two novel methods in this line of research: eye-tracking and linear mixed effects modelling.

The paper reports on two studies. The first study (Holzknecht et al., 2017) looked at cognitive processing in listening items on the Aptis General Test using eye-tracking and stimulated recall. The study did not focus on response order. In fact, we used it as a control variable, but we found a strong and unexpected effect of order in the eye-tracking data. In the current paper, we are recasting response order as the main variable under investigation. The second study, which led on from the first, looked at the direct effect of response order on item difficulty on 200 live Aptis listening items.

Methods

Regression models. Both studies presented in this paper use multiple regression models to respond to the goals of the research. For a non-technical yet comprehensive explanation of these kinds of models see, for example, Field (2013). Study 2 uses a standard multiple regression model, whereas Study 1 uses a slightly more complex linear mixed effects multiple regression model. However, fundamentally these two models are very similar in that they allow us to examine the association between one outcome variable (i.e., the dependent variable in which we are particularly interested) and one or more explanatory variables which are used to explain the observed values of the dependent variable.

Whereas correlation coefficients only allow us to examine the relationship between two variables, multiple regression looks at the relationships between multiple variables simultaneously in order to investigate one variable of interest. This is particularly useful when we want to control for (i.e., statistically take into consideration) the effect of specific variables. For example, we may see a correlation between *L2 ability* and *age* in a dataset, leading us to conclude that there is a positive linear association between these two variables. However, if we simultaneously model *age* and *number of years of L2 study* in a multiple regression model predicting *L2 ability* (i.e., controlling for *number of years of L2 study*), the association between *age* and *L2 ability* may well fall away, as it was simply owing to the fact that older people have had a longer time in which to learn a language. In this example, controlling for one variable (i.e., *number of years of L2 study*) allowed us to more clearly see the true relationship between another pair of variables (i.e., *L2 ability* and *age*). In the studies outlined below, we needed to control for multiple factors to investigate the true relationships with regard to response order effects.

Study 1: Eye-tracking study of correct response location

Participants. A total of 30 participants (14 male, 16 female) took part in this study. The participants were all native speakers of German and were aged between 20 and 61 years, with a mean age of 28.5 years. All participants completed three Aptis General components (grammar and vocabulary, listening, and reading,) and attained high average scores out of a maximum of 50 points: grammar and vocabulary ($M = 36.83$, $SD = 6.77$), listening ($M = 42.00$, $SD = 5.50$), and reading ($M = 45.50$, $SD = 5.75$).

Materials. A retired Aptis General listening module consisting of 25 MC items was used in the data collection. All 25 MC items had a question, a stem, and four options from which to choose, with one correct option and three distractors for each item. The test included several items from four CEFR levels: A1 ($n = 7$ items), A2 ($n = 7$ items), B1 ($n = 6$ items), and B2 ($n = 5$ items). The items were presented starting at A1 level and then increased in difficulty, as is also the case for the operational Aptis Test. Thus, each participant saw the items in the exact same sequence.

The stimuli were created as html files using Verdana (font size 32px/24pt) on a 23-inch monitor (1920 × 1080) and then integrated into the Tobii Studio eye-tracking software. The original layout of the Aptis Test was slightly altered in that the response options were moved further apart by including more blank space between the options (see Figure 1). This step helped improve data quality as areas of interest could be specified more clearly. The eye-tracking was conducted using a Tobii TX300 (300Hz sampling rate, accuracy 0.4°). All participants sat approximately 63cm from the monitor and the distance from screen was monitored throughout the experiment in the "Track Status" window of Tobii Studio (see also the detailed description of the procedure below). The experiment consisted of eight sets of three-to-four items each, with a total of 25 items for each candidate. None of the items required scrolling. The Tobii I-VT filter was used, with a velocity threshold of 30 degrees per second, a window length of 20ms, and a minimum fixation duration of 60ms (see Olson, 2012 for the rationale of these values for the collection of reading data). Velocity threshold (VT) filters are considered particularly

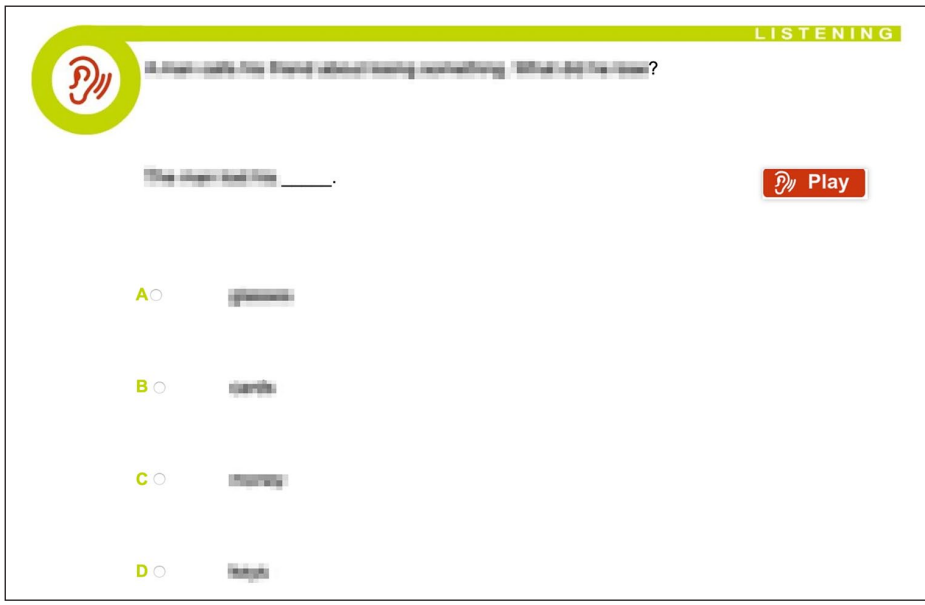

Figure 1. Example stimulus item.
Note: Item, stem, and response options have been blurred for reasons of test security.

suitable for the analysis of reading data from high-speed eye-trackers (Holmquist et al., 2011) such as the one used in this experiment.

Procedure. After obtaining ethical approval from the ethics committee at the university of the researchers, participant recruitment commenced. All participants received written information outlining the study and signed consent forms. Before the experiment, participants were shown a sample MC item and were instructed on how to start the sound file and respond to items. Participants were also told that they could decide themselves whether or not they wanted to listen to a sound file once or twice. This is consistent with the operational Aptis General Test. Participants were reminded to answer the items as if they were taking an actual language test. The instructions were provided in the participants' L1 and they were encouraged to ask questions if something seemed unclear.

After these initial instructions, each participant also received instructions and explanations concerning the eye-tracking. This included adjusting the participant's seating position and a short demonstration on how moving the body or head impacts eye-tracking quality. In a last step before commencing data collection, participants were instructed to place the index finger of their left hand on the ESC key to be able to move on from one item to the next, and their right hand onto the mouse. These instructions were adapted for left-handed participants and were meant to help all participants navigate the test items without looking off-screen.

The experiment started with a standard five-point calibration of the Tobii TX300 eye-tracker and finding a comfortable and optimal seating position (i.e. approximately 63cm

from screen, as measured by the Tobii Studio software) for the participant. Calibration was repeated until reaching a satisfactory level of accuracy before starting the series of experiments. The eye-tracker was then recalibrated at the beginning of each set of items which helped ensure accuracy throughout the data collection (see also Conklin, Pellicer-Sánchez, & Carrol, 2018 for practical recommendations for setting up eye-tracking experiments). The participants were asked to remain as still as possible after calibration. Then the first set of three items was presented. Throughout each set of items, the researchers monitored the head position and eye-tracking quality via the "Track Status" function of Tobii Studio and provided feedback in between the items in case a participant changed their position. However, as the Tobii TX300 allows for some natural movement of the head this was not necessary with most participants. Great care was taken to avoid discomfort or strain for the participants while working on the items and they were given shorter breaks upon completing a set of items and a longer break after the first five sets (i.e., after the first 15 items).

Measures. While there are no additional stimuli such as pictures or graphs presented in Aptis Listening Test items, we still expected that eye-gaze patterns on the textual information of the items could reveal insights into participants' test-taking behaviour. To be able to test hypotheses related to eye gazes, areas of interest were defined for each aspect of the stimuli relevant to the research aims. Each of the four response options was defined as a separate area of interest. The dependent variable for subsequent analyses then was the total visit duration on each of the four options. Total time is defined as the summed duration of all visits by each participant and on each area of interest and needs to be understood as a global measure as it aggregates all gaze activity, such as first fixation durations as well as any re-reading activities within an area of interest (Godfroid, 2019). It is thus useful in assessing global effects such as comparing length of processing for each of the four response options. The hypothesis to be tested was whether total visit duration was the same for each of the four options, after controlling for a number of variables as outlined in the following.

A regression model was used to allow a multivariate analysis of the total visit durations for each response, on each item, and by each person. As outlined above, the advantage of using multiple regression models is that all variables can be modelled jointly, without having to use average values across variables and while still being able to control for different variables. Table 1 below outlines all variables included in the analysis. The three control variables added to the model were the response chosen (participants were found to focus naturally more on their final choice, as they had to move the mouse towards their chosen option), the number of times they listened to the sound file (participants could choose between listening once or twice as is operational in the Aptis Listening Test and this impacted the time available for looking at the options), and the CEFR level of the item (items at different levels were developed with the intention to tap into different cognitive processing). Not controlling for these three variables might run the risk of drawing wrong conclusions on the basis of confounded variables.

Statistical analysis. For the data analysis, a mixed effects linear regression model (Gelman & Hill, 2006) including random intercepts was chosen. This method of analysis

Table 1. Measures used in Study 1.

Variable name	Technical description	Reason for including
Visit duration	The visit duration, measured in seconds, for a particular individual on a particular item on a particular response. (*Dependent variable*)	This measure represents the amount of processing directed towards a response option.
Response order	The order in which the responses were presented on the screen (*Variable under investigation*)	This variable was the variable under investigation.
Participant	A factor indicating which participant the visit duration came from (*Random effect*)	This variable was included to model and control for individual differences.
Item	A factor indicating which item the visit duration came from (*Random effect*)	This variable was included to model and control for differences between items.
Response chosen	A binary indicator of whether the particular response option was chosen by the participant (*Control variable*)	This variable controlled for the fact that participants put additional focus on the option chosen owing to (1) the processing which occurs when matching the text to the representation of the chosen answer, and (2) the need to execute fine motor control with visual feedback to click the mouse in the correct location.
Listen times	The number of times the participant listened to the text of a particular item (once or twice) (*Control variable*)	In the Aptis Listening Test, participants can choose whether to listen to the text of an item once only or twice. Therefore, the number of listening times needed to be controlled for in the analysis before comparing visit durations between items of different CEFR levels.
CEFR level	The British Council assigned CEFR level of the item (*Control variable*)	This variable controls for the fact that more cognitively complex items should require more processing on the response options to answer.

has been used successfully in other linguistics research projects and has been recommended by various researchers (Baayen, Davidson, & Bates, 2008; Winter, 2013). In mixed effects models, the predictor variables can be classified as either fixed or random, whereby the fixed parameters are the factors under investigation and the random variables come from a whole range of potential parameters.

As listed in Table 1, the fixed variables for this study were the three control variables (response chosen, audio replayed or played once, and CEFR level) and the variable under investigation (response order). These four factors were considered relevant to the processing of an individual test taker when answering particular items and response options. The random effects were participant and item. The goal of this study was to investigate

Table 2. Measures used in Study 2.

Variable name	Technical description	Reason for including
Logit difficulty	The item difficulty value of the specific items as measured in logits.	This is the dependent variable. As we are interested in assessing the extent to which difficulty is influenced by position on screen, we need to include this measure.
CEFR level	The CEFR level particular items were written and accepted after extensive piloting to be examining.	This variable is used to correct for imbalances in the position of the correct location across items types.
Correct response position	The location of the correct response for a particular item.	This is the main explanatory variable. We are interested in discovering whether this variable affects items difficulty.

the effect of response order to be able to generalize from the sample to a broader population of items and participants. However, certain individuals or certain items may be more prone to produce a certain visit duration and be correlated. The strength of the mixed effect model approach is that such correlations can be taken into account by including random effects in the model. Not including random parameters in the model can distort results and lead to invalid inferences about the statistical significance of the fixed effects (Crawley, 2007).

The regression analysis was carried out with the package *lme4* (version 1.1-21; Bates, Maechler, Bolker, & Walker, 2015) for *R* (version 3.5.1; R Core Team, 2014). The random effects, participant and item, were characterized by a random intercept. The *p*-values for the fixed effects were obtained via Satterthwaite approximation using the package *lmerTest* (version 3.1-1; Kuznetsova, Brockhoff, & Christensen, 2016).

Study 2: Investigation of item difficulty based on correct response position

Materials. The materials for this study represent psychometric characteristics of 200 live Aptis listening comprehension items. In terms of CEFR level, 40 of these items were aimed at A1, 52 at A2, 50 at B1 and 58 at B2. The psychometric properties of the items were based on a total sample of between 5821 and 6166 test takers, differing slightly for individual items.

Measures. Three measures were modelled and descriptions of these measures are shown in Table 2. The dependent variable *logit difficulty* is a continuous variable representing the item difficulty for a particular item, expressed on the logit scale. The variable *CEFR level* is a categorical variable from A1 to B2 used to adjust for any possible imbalance in correct response location. For example, if the A1 items had more items with the correct response in the first position than the B2 items, spurious conclusions about the effect of correct answer position on difficulty could arise. The explanatory variable "correct response position" is a categorical variable ranging from first to fourth which specifies the location of the correct option on the page in the live Aptis Tests. The true positions of the correct answers are roughly evenly distributed across the four

possibilities and we are assured by the test developers were selected at random, while maintaining approximate equality in correct answer location in an individual form of the test. It should be noted that the Aptis Test does not change the location of the correct response for a specific item.

Statistical analysis. A linear regression model, fit via Ordinary Least Squares, was used to estimate parameters. The function *lm* in *R* (version 3.5.1; R Core Team, 2014) was used to fit the model.

Results

Study 1: Eye-tracking study of correct response location

Visit duration, the dependent variable, emerged as highly negatively skewed. Therefore, we decided to *log* transform the raw values. This approach was preferred over using a Gamma distribution in order to increase the interpretability of the regression model, as coefficients based on *log*-transformed data can be understood simply in terms of percentage increase or decrease of the variable, in this case visit duration, under the different conditions. Although the *log*-transformed data was still skewed, it did fit the main assumption of regression in that the model's residuals were normally distributed. Once the model had been fitted, residual plots were used to confirm that the assumptions of homoscedasticity and normality had not been violated. One outlier had to be removed owing to its residual.

Table 3 shows the results of the linear mixed model. Considering model parsimony, including both random effects, participants and items, was found more useful than including none or just one (Akaike, 1974). While some of the total variance is explained by the two random effect variables, the majority of variance remains unexplained as a residual. Table 3 also reports the raw β estimate with its standard error (*SE*), the approximate degrees of freedom (Approx. *df*) and significance value (*p*-value) for each of the predictors included in the model. The results indicate that the model explains a substantial portion of the variance in visit durations with high squared correlation between observed and fitted values ($R^2 = 0.68$, bottom of table).

Without *log* transformation of the dependent variable, there are usually two ways in which β estimates of categorical and continuous variables can be interpreted. In the case of categorical variables, the estimate stands for the amount to be subtracted (or added) from the intercept when the data point is located in that particular category, and for continuous variables the β estimate represents a 1-unit decrease (or increase) in the dependent variable. With *log* transformation of the dependent variable (as was the case in this study), the interpretation of the β estimate needs to be slightly adapted. The exponential function (exp (β estimate)) inverses the natural logarithm and can be translated into the effect size of being part of that category for categorical data or, in case of a continuous variable, into a percentage decrease or increase for a 1-unit increase in the explanatory variable. The percentage decreases and increases in terms of expected visit duration are given in parentheses after the β estimate in Table 3. The exponential of the intercept (4.42 seconds) is the fitted total visit duration on a particular item's response option, with the following characteristics:

Table 3. Results of linear mixed model for Study 1.

Random effects	Variance	*SD*		
Participants	0.08	0.28		
Items	0.08	0.28		
Residual	0.39	0.62		
Fixed effects	**β estimate**	***SE***	**Approx. *df***	***p*-value**
(intercept)	1.49 (4.42s)	0.12	32	0.00***
Response order 2	**–0.38 (–30%)**	**0.03**	**2913**	**0.00*** **
Response order 3	**–0.82 (–56%)**	**0.03**	**2914**	**0.00*** **
Response order 4	**–1.39 (–75%)**	**0.03**	**2914**	**0.00*** **
Listening twice	0.09 (+9%)	0.03	2939	0.01**
Chosen response	0.79 (+120%)	0.03	2914	0.00***
CEFR A2	0.34 (+40%)	0.15	21	0.04*
CEFR B1	0.90 (+145%)	0.16	21	0.00***
CEFR B2	1.22 (+240%)	0.17	22	0.00***

Squared correlation between observed and fitted (pseudo R^2) = 0.68.

- The item was at A1 level.
- The response option was first on the page.
- The response option was not chosen by the participant.
- The participant listened to the text once.
- The participant had listening, reading, and eye-tracking test scores of 0.

In the following, the results on the different fixed effects included in the model as displayed in Table 3 will be described.

Response order. As can be seen in Table 3, the position in which a response option is presented to a participant plays a clear and highly statistically significant role: the further up a response on the page the longer the participant would look at this particular response and the lower down the response, the shorter the visit duration. Relative to the first option, the total visit duration decreased by 30% for the second option, by 56% for the third option and by 75% for the fourth option.

Listening twice. There is a statistically significant effect of listening twice to the sound file in that visit duration on the response options increases by 9%. As this effect appears surprisingly small, it might be the case that this variable correlates with the CEFR level. In other words, higher CEFR-level items tended to generate more instances where a candidate listened to the sound file twice, meaning that the variance of this variable is mainly explained by the CEFR level. Nonetheless, it is important to include this variable as a predictor to control for the fact that whether or not candidates listened once or twice was not regulated by the experimental design.

Table 4. Results of multiple regression model for Study 2.

	β estimate (logits)	*SE*	*t*-value	*p*-value
(intercept)	–2.69	0.20	–13.25	0.00***
Response order 2	**0.27**	**0.20**	**1.32**	**0.19**
Response order 3	**0.29**	**0.20**	**1.50**	**0.13**
Response order 4	**0.48**	**0.22**	**2.15**	**0.03***
CEFR A2	1.14	0.22	5.26	0.00***
CEFR B1	2.27	0.22	1.39	0.00***
CEFR B2	2.77	0.21	13.04	0.00***

Adjusted $R^2 = 0.51$.
F-statistic = 35.92 ($df = 6$ & 193), $p = 0.00$***.

Chosen response. The response that participants finally opted for during the experiment affected the total visit duration of that response by an increase of 120%. This result is not surprising for two reasons. First, participants are likely to have spent more time looking at an option that they finally selected because of increased processing of this option and confirming their choice. Second, participants had to select manually the option via mouse click, which also increases the time they spend near the option to perform the clicking correctly. As with the variable of listening twice, this variable was included into the model, as it cannot be controlled for in the experiment and could lead to confounding with response order, for reasons discussed above, and thus with the CEFR level.

Study 2: Investigation of item difficulty based on correct response position

Table 4 shows the results of a linear regression model explaining item difficulty, as measured in logits, with the intended CEFR level of the item and the location of the correct response on the page. As shown at the bottom of the table, a large amount of variance is explained by this model (adj $R^2 = 0.51$).

The model suggests that responses in the second, third and fourth positions on screen are 0.27, 0.29 and 0.48 logits more difficult than those in the first position, respectively. Although there is only one statistically significant difference between the categories in the response order variable, (i.e. the difference between the first and fourth), it should be noted that (1) the difficulties are in line with the hypothesis and (2) this analysis is only based on a relatively small number of items and is thus unlikely to find significance for what are relatively small but practically important effect sizes.

Discussion

Given the widespread use of MC tasks in language assessment programs around the world (both in low-stakes and high-stakes situations), potential test method effects on item difficulty need to be investigated thoroughly. The current study adds to the body of research on effect of response order on item difficulty in MC tests, as previous findings

have been inconclusive. It did so by looking at such effects in relation to L2 listening assessment (a skill not yet researched in this area), using two new methods in this line of research: eye-tracking and linear mixed effects modelling.

In the first part of the study, we showed, through linear mixed effects modelling of eye-tracking data, that when solving items on the Aptis Listening Test, participants focused significantly longer at responses higher up the screen, with a clear progression from the top to the bottom of the screen. Test takers looked at the first responses 30% longer than at the second, 56% longer than at the third, and 75% longer than at the fourth. To our knowledge, it is the first time that such an effect is shown in relation to MC testing. One factor that could have contributed to these results is that test takers may not read subsequent options if they identify the correct option in the earlier response positions. This is also a commonly taught test taking strategy and especially useful for tests where time is a concern. However, although this phenomenon seems consistent with the primacy effect hypothesis found in research across disciplines, it does not explain whether item difficulty is affected by it.

In order to test whether items are more difficult when the key is placed in later positions, responses on 200 Aptis Listening Test items by about 6000 candidates for each item were modelled with regards to key position and item difficulty. The regression analysis showed that items with the key in fourth position are significantly more difficult than items with the key in first position, with a difference of 0.48 logits. These results confirm findings by Sonnleitner et al. (2016), who reported the same effect for MC vocabulary items. However, the findings from the present study might be taken to be a more appropriate reflection of a test-taking candidature as the Sonnleitner et al. (2016) study was based on a less authentic experimental setup. Our findings are also in line with results from Hohensinn and Baghaei (2017), who found the same tendency for a test consisting of grammar, vocabulary, and reading items, but argued that the effect was only very small and not of practical significance. However, Hohensinn and Baghaei (2017) only looked at a total of 60 four-option MC items across three different competency areas, so statistically significant results may not have emerged owing to the small sample size. Our sample of items was considerably larger (200) and more homogeneous as it only consisted of listening items.

These results are of importance for the development of MC listening tests, for three main reasons. First, test developers need to consider this primacy effect when deciding on the number of options. Other research has highlighted the advantages of three options compared to four options or more. For example, Haladyna and colleagues have shown that MC items with four options mostly have only one or two distractors with acceptable discrimination (Haladyna & Downing, 1993), and that developing a third distractor is also difficult from an item writer perspective (Haladyna, Downing, & Rodriguez, 2002). In addition, Rodriguez (2005) argued that more items can be administered when using only three options as test takers need less time to read all options, which in turn can be beneficial for reliability and construct representation. Our study supports the argument for three options by indicating that for four-option MC listening items ordering effects impact item difficulty, but that this effect is less pronounced when only three options are used.

Second, the findings show that in the item writing process careful consideration needs to be given to the positioning of the key, as this might influence item difficulty. Our results indicate that MC listening tests with many keys in later positions would likely lead to fewer correct answers than tests with many keys in earlier positions. This effect should thus be taken into account when comparing item difficulty across tests but also across different versions of the same test.

Lastly and most importantly, the findings question the commonplace practice of randomizing correct answer position for each individual candidate. Any given item may therefore vary in difficulty for individual candidates. This is particularly pertinent as it is currently unclear whether the effect is the same across items targeting different proficiency levels. For example, it might be the case that the effect is more pronounced in more difficult items. Since any item in this scenario has four different difficulty levels depending on the correct response position, randomization would introduce construct irrelevant variance. This then makes this practice an issue of fairness, as randomization would mean that for some candidates the same test would be more difficult than for others.

Future research

Future research should continue to explore this issue in MC listening assessment using controlled experimental design with larger sample sizes. Following Sonnleitner et al.'s (2016) research design, researchers could manipulate the position of the correct answer to counterbalance answer positions across participants and items. This could be done in both computerized and paper-and-pencil tests to probe whether a response order effect can be detected regardless of delivery mode, or whether different modes are affected differentially. In addition, such experimental designs might find it useful to control for a number of comprehension behaviours targeted in items to investigate whether, for instance, items eliciting comprehension of specific details are affected differently than items assessing inferencing. Moreover, it may be relevant to examine the effect in variations of MC items (three, four, five, or more options).

In addition, it would also be of interest to investigate potential explanatory variables of response order bias. For example, one could hypothesize that a primacy effect might be stronger in impulsive test takers. Using standardized questionnaires to establish test takers' levels of impulsiveness (e.g., the Barratt Impulsiveness Scale, Barratt, 1994) would thus shed further light on test takers' response processes and help to explain what person-level factor may cause not reading all the response options equally. Finally, future studies could further investigate potential primacy effects across different skills, candidates of different proficiency levels, and items targeting different proficiency levels.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Part of this research was funded by the British Council through the Assessment Research Awards and Grants scheme (AR-G/2017/3).

ORCID iD

Franz Holzknecht https://orcid.org/0000-0002-1218-2062

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/tac.1974.1100705

Andringa, S., Olsthoorn, N., Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, *62* (Supplement s2), 49–78. https://doi.org/10.1111/j.1467-9922.2012.00706.x

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, *41*, 258–290. https://doi.org/10.1037/h0055756

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109–128. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x

Baayen, R. H., Davidson, D. J., & Bates, D. (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

Ballard, L. (2017). The effects of primacy on rater cognition: An eye-tracking study. Unpublished PhD dissertation. Michigan State University. Michigan, US. https://d.lib.msu.edu/etd/4632/datastream/OBJ/download/THE_EFFECTS_OF_PRIMACY_ON_RATER_COGNITION___AN_EYE-TRACKING_STUDY.pdf

Barratt, E. S. (1994). Impulsiveness and aggression. In J. Monahan & H. J. Steadman (Eds.), *Violence and mental disorder: Developments in risk assessment* (pp. 61–79). University of Chicago Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bolhuis, J. J., & Bateson, P. (1990). On the importance of being first: A primacy effect in filial imprinting. *Animal Behavior*, *40*, 472–483. https://doi.org/10.1016/s0003-3472(05)80527-5

Bruine de Bruin, W. (2005). Save the last dance for me: unwanted serial position effects in jury evaluations. *Acta Psychologica*, *118*(3), 245–260. https://doi.org/10.1016/j.actpsy.2004.08.005

Brunfaut, T. (2016). Assessing listening. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 97-112). (Handbooks of Applied Linguistics; Vol. 12). Boston/Berlin: Mouton de Gruyter. https://doi.org/10.1515/9781614513827-009

Brunfaut, T., & Révész, A. (2015). The role of listener- and task-characteristics in second language listening. *TESOL Quarterly*, *49*(1), 141–168. https://doi.org/10.1002/tesq.168

Buck, G. (2001). *Assessing listening*. Cambridge University Press. https://doi.org/10.1017/cbo9780511732959

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*, 323–331. https://doi.org/10.1016/j.jarmac.2018.07.002

Carney, D. R., & Banaji, M. R. (2012). First is best. *PLoS ONE*, *7*(6), 1–5. https://doi.org/10.1371/journal.pone.0035088

Carrell, P. L. (2007). Notetaking strategies and their relationship to performance on listening comprehension and communicative assessment tasks. *TOEFL Monograph Series* RR-07-01. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02043.x

Christenfeld, N. (1995). Choices from identical options. *Psychological Science*, *6*(1), 50–55. https://doi.org/10.1111/j.1467-9280.1995.tb00304.x

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement*, *54*, 8–20. https://doi.org/10.1177/0013164494054001002

Clark, E. L. (1956). General response patterns to five-choice items. *Journal of Educational Psychology*, *47*(2), 110–117. https://doi.org/10.1037/h0043113

Conklin, A., Pellicer-Sánchez, K., & Carrol, G. (2018). *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press. https://doi.org/10.1017/9781108233279

Crawley, M. J. (2007). Mixed-effects models. In M. J. Crawley (Ed.), *The R Book, Second Edition* (pp. 681–714). John Wiley and Sons.

Dayan, E., & Bar-Hillel, M. (2011). Nudge to nobesity II: Menu positions influence food orders. *Judgment and Decision Making*, *6*(4), 333–342. http://journal.sjdm.org/11/11407/jdm11407.html

Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *The Modern Language Journal*, *89*(2), 206–220. https://doi.org/10.1111/j.1540-4781.2005.00275.x

Ert, E., & Fleischer, A. (2016). Mere position effect in booking hotels online. *Journal of Travel Research*, *55*(3), 311–321. https://doi.org/10.1177/0047287514559035

Fagley, N. S. (1987). Positional response bias in multiple-choice tests of learning: Its relation to testwiseness and guessing strategy. *Journal of Educational Psychology*, *79*(1), 95–97. https://doi.org/10.1037/0022-0663.79.1.95

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). SAGE Publications. https://doi.org/10.1024/1012-5302/a000397

Field, J. (2015). The effects of single and double play upon test outcomes and cognitive processing. *ARAGs Research Reports Online*. The British Council. www.britishcouncil.org/sites/default/files/field_layout.pdf

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Godfroid, A. (2019). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge. https://doi.org/10.4324/9781315775616

Green, R. (2017). *Designing listening tests: A practical approach*. Palgrave Macmillan. https://doi.org/10.1057/978-1-349-68771-8

Greenwald, A. G., Pickrell, J. E., & Farnham, S. D. (2002). Implicit partisanship: Taking sides for no reason. *Journal of Personality and Social Psychology*, *83*(2), 367–379. https://doi.org/10.1037/0022-3514.83.2.367

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice item? *Educational and Psychological Measurement*, *53*, 999–1010. https://doi.org/10.1177/0013164493053004013

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*, 309–334. https://doi.org/10.1207/S15324818AME1503_5

Hohensinn, C., & Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicologica*, *38*, 93–109. http://files.eric.ed.gov/fulltext/EJ1125979.pdf

Holmquist, K., Nyström, N., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (Eds.). (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Holzknecht, F. (2019). Double play in listening assessment. Unpublished PhD dissertation. Lancaster University. Lancaster, UK.

Holzknecht, F., Eberharter, K., Kremmel, B., McCray, G., Zehentner, M., Konrad, E., & Spöttl, C. (2017). *Looking into listening: Using eye-tracking to establish the cognitive validity of the Aptis Listening Test*. ARAGs Research Reports Online. The British Council.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244. https://doi.org/10.1177/0265532208101006

Insko, C. A. (1954). Primacy versus recency in persuasion as a function of the timing of arguments and measures. *Journal of Abnormal and Social Psychology*, *69*, 381–391. https://doi.org/10.1037/h0042765

Jersild, A. (1928). Modes of emphasis in public speaking. *Journal of Applied Psychology*, *12*, 611–620. https://doi.org/10.1037/h0075226

Johnson, M. H. (1992). Imprinting and the development of face recognition: From chick to man. *Current Directions in Psychological Science*, *1*, 52–55. https://doi.org/10.1111/1467-8721.ep11509740

Jones, E. E., Rock, L., Shaver, K. G., Goethals, G. R., & Ward, L. M. (1968). Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology*, *10*, 317–347. https://doi.org/10.1037/h0026818

Knower, F. H. (1936). Experimental studies of change in attitude: II. A study of the effect of printed argument on changes in attitude. *Journal of Abnormal and Social Psychology*, 522–532. https://doi.org/10.1037/h0055902

Kormos, J., & Sáfár, A. (2008). Phonological short term-memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, *11*(2), 261–271. https://doi.org/10.1017/s1366728908003416

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, *51*, 201–219. https://doi.org/10.1086/269029

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in linear mixed effects models. R package version 2.0-3.0*. http://cran.r-project.org/package=lmerTest

Macaro, E., Vanderplank, R., & Graham, S. (2005). *A systematic review of the role of prior knowledge in unidirectional listening comprehension*. EPPI-centre. http://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/MFL_rv2.pdf?ver=2006-03-02-125000-953

MacIntyre, P. D., & Gardner, R. C. (1991). Methods and results in the study of anxiety and language learning: A review of the literature. *Language Learning*, *41*(1): 85–117. https://doi.org/10.1111/j.1467-1770.1991.tb00677.x

Marcus, A. (1963). The effect of correct response location on the difficulty level of multiple-choice questions. *Journal of Applied Psychology*, *47*(1), 48–51. https://doi.org/10.1037/h0042018

Miller, M., & Campbell, D. T. (1959). Recency and primacy in persuasion as a function of the timing of speeches and measurements. *Journal of Abnormal and Social Psychology*, *59*, 1–9. https://doi.org/10.1037/h0049330

Olson, A. (2012). The Tobii I-VT fixation filter. https://www.tobiipro.com/siteassets/tobii-pro/learn-and-support/analyze/how-do-we-classify-eye-movements/tobii-pro-i-vt-fixation-filter.pdf

Pineno, O., & Miller, R. R. (2005). Primacy and recency effects in extinction and latent inhibition: A selective review with implications for models of learning. *Behavioural Processes*, *69*, 223–235. https://doi.org/10.1016/j.beproc.2005.02.006

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. www.r-project.org

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*(1), 31–65. https://doi.org/10.1017/s0272263112000678

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Rosenhouse, J., Haik, L., & Kishon-Rabin, L. (2006). Speech perception in adverse listening conditions in Arabic-Hebrew bilinguals. *International Journal of Bilingualism*, *10*(2), 119–135. https://doi.org/10.1177/13670069060100020101

Sonnleitner, P., Guill, K., & Hohensinn, C. (2016). *Effects of correct answer position on multiple-choice item difficulty in educational settings: Where would you go?* Paper presented at the International Test Commission Conference, Vancouver, Canada.

Strauss, D. J., & Francis, A. L. (2017). Toward a taxonomic model of attention in effortful listening. *Cognitive, Affective and Behavioral Neuroscience*, *17*(4), 809–825. https://doi.org/10.3758/s13415-017-0513-0

Taylor, A. K. (2005). Violating conventional wisdom in multiple choice test construction. *College Student Journal*, *39*, 141–153.

Tellinghuisen, J., & Sulikowski, M. M. (2008). Does the answer order matter on multiple-choice exams? *Journal of Chemical Education*, *85*(4), 572. https://doi.org/10.1021/ed085p572

Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics*, *26*(1), 70–89. https://doi.org/10.1093/applin/amh039

Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency? *The Modern Language Journal*, *90*(1), 6–18. https://doi.org/10.1111/j.1540-4781.2006.00381.x

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge. https://doi.org/10.4324/9780203843376

Winke, P., & Lim, H. (2014). The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation. *IELTS Research Reports Online Series*, *3*, 1–30. https://www.ielts.org/-/media/research-reports/ielts_online_rr_2014-3.ashx

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 37–53. https://doi.org/10.1016/j.asw.2015.05.002

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. ArXiv:1308.5499. https://arxiv.org/abs/1308.5499