| | |
|---|---|
| Title | An eye-tracking investigation of the keyword-matching strategy in listening assessment |
| Author(s) | Shermaine Qi En Kho, Vahid Aryadoust and Stacy Foo |

# An Eye-Tracking Investigation of the Keyword-Matching Strategy in Listening Assessment

Citation:

Kho, S. Q. E., Aryadoust, V., & Foo, S. (2022). An eye-tracking investigation of the keyword-matching strategy in listening assessment. *Education and Information Technologies*. Advance online publication. https://link.springer.com/article/10.1007/s10639-022-11322-y

*Shermaine Qi En Kho*
National Institute of Education
Nanyang Technological University
Singapore
ORCID: https://orcid.org/0000-0002-2706-726X
shermainekho19@gmail.com

*Vahid Aryadoust * (corresponding author)*
National Institute of Education
Nanyang Technological University
Singapore
Vahid.aryadoust@nie.edu.sg

*Stacy Foo*
National Institute of Education
Nanyang Technological University
Singapore
stacy.w.l.foo@gmail.com

**Conflict of Interest**: None

## ACKNOWLEDGMENT

We would like to acknowledge the primary funding support from Paragon Testing Enterprises, Canada. The secondary funding support for this project came from Nanyang Technological University (Singapore) under the Undergraduate Research Experience on CAmpus (URECA) programme.

## Abstract

Studies have shown that test-takers tend to use keyword-matching strategies when taking listening tests. Keyword-matching involves matching content words in the written modality (test items) against those heard in the audio text. However, no research has investigated the effect of such keywords in listening tests, or the impact of gazing upon these keywords on listening test scores. Thus, this study examined whether test-takers' performance on a listening test can be explained by their gaze behaviors across three types of content words in the written modality: nouns, verbs, and adjectives. Using eye-tracking technology, this study measured the gaze behavior of 66 listening test-takers during reading content words in test item stems. Using linear mixed effect model, binary probit regression, and multinomial logistic regression, we found that test-takers' performance was predicted by gaze behavioral measures on content words. Among the content words, fixating on nouns in written test items had the most significant role in predicting test performance, followed by adjectives, and verbs. By shedding light on how keywords in test items are attended to by test-takers and the relationship between keyword-matching and listening test performance, this study has provided significant evidence for the overwhelming role of reading in listening tests. Implications for test score interpretation are discussed.



***Keywords*:** eye-tracking; gaze behaviors; keyword-matching; listening comprehension tests; multimodality.


# An Eye-Tracking Investigation of the Keyword-Matching Strategy in Listening Assessment

An extensive body of empirical research has aimed to explain the cognitive processes underlying listening comprehension, commonly referred to as bottom-up and top-down processing (Bodie et al., 2020; Kintsch, 1998; Masrai, 2020; Nation & Newton, 2008; Tsui & Fullilove, 1998; Wilson, 2003; Wu, 1998). In bottom-up processing, listeners focus on the literal and linguistic meanings, deciphering information from the "auditory-phonetic, phonemic, syllabic, lexical, syntactic, semantic, propositional, pragmatic and interpretive" to make sense of the listening input (Nation & Newton, 2008, p. 40). Thus, this mode of listening emphasizes single-word recognition, among other things, corresponding to the listener's existing lexicon, employing "phonological, lexical, syntactic and semantic knowledge" as tools in decoding the listening input (Wu, 1998, p. 22). Conversely, top-down processing requires listeners to employ inferential skills to extract the gist of the listening input (Aryadoust, 2020; Dunkel et al., 1993). Here, listeners integrate newly encountered information with pre-existing or background knowledge to resolve ambiguities in the listening input (Goh, 2002; Nation & Newton, 2008; Wu, 1998). Such integration requires listeners to draw associations among elements of speech in recently discerned parts of the listening input with previously processed information so as to formulate a coherent interpretation of the input (Daneman & Merikle, 1996).

Understanding what cognitive processes contribute to listening performance is especially important when considering the types of listening assessments (e.g., Wagner, 2010). From the test methods perspective, listening assessments can be divided into while-listening performance and post-listening performance tests (Aryadoust, 2011, 2019). A while-listening performance test requires test-takers to simultaneously listen to the audio texts, read and answer the test items. In contrast, a

post-listening performance test demands test-takers to first listen and take notes of the listening input, and then answer the test items Aryadoust, 2011, 2018).

Although while-listening performance tests are used in the listening sections of many international standardized tests of English language proficiency, such as the Canadian Academic English Assessment (CAEL) and the International English Language Testing System (IELTS), they are limited in several ways. These tests require test-takers to process continuous incoming information from the listening input while answering questions concurrently. Such multitasking could increase the cognitive load of test-takers, resulting not only in the splitting of attention but also overloading the working memory, thus challenging test-takers (Aryadoust, 2011, 2019; Wickens, 2006). Consequently, test-takers use keyword-matching to ease cognitive load (Field, 2009), which in turn influences test performance in different ways, like listening test-takers' dividing their working memory capacity between auditory and written modalities. Fixating on keywords in the written modality and matching them against the auditory modality seems to represent an inauthentic reading-and-listening process that is directed by words in the written modality. This cognitive process is unlike what cognitive theories of listening comprehension describe as listening (Field, 2013). As a result, the test scores that represent this cognitive process would not represent the actual listening abilities of test-takers.

While-listening performance procedures can be contrasted with the retrospective procedures in post-listening performance tests which make the working memory available without interfering with processing, and are thus less taxing on processing capacities (Field, 2009). It is also known that test-takers find reading test items in the absence of the listening stimuli less challenging, as they could focus

entirely on reading the test items instead of dividing their attention across written and auditory sources of information (Haarmann et al., 2003). Due to the problematic operationalization of listening comprehension constructs (Field, 2009), while-listening performance tests may not discriminate between listening comprehension skills and the working memory available without interfering with processing.

In addition, as the developers of some these tests have claimed, the design of items in these tests focuses on "explicit and easily accessible information" (Geranpayeh and Taylor, 2008) rather than a coherent interpretation of the input, thus compromising the cognitive validity and authenticity of the tests by testing comprehension only on a superficial level and encouraging shallow listening (see Field, 2009). The above-mentioned findings and claims have sparked investigation into the cognitive validity of listening tests, that is, whether skills tested in listening tests resemble those involved in natural listening and if they really do reflect the actual listening proficiency of the test-taker, given the artificial setting of a test (Badger & Yan, 2009; Field, 2013). Geranpayeh and Taylor (2008) discussed the complex interplay of cognitive and contextual factors, especially in deciding whether or not the listening input can be replayed under test conditions, thus questioning the authenticity and cognitive validity of such listening tests (see Douglas, 2001; Ockey & Wagner, 2018; Ryan & Granville, 2020; Weir, 2005; and Wood, 1993, for a review of these concepts). Meanwhile, Goh (2002) and Maftoon and Alamdari (2020) reported that test-takers and language learners use a multitude of cognitive and metacognitive tactics and strategies in listening comprehension tests such as focusing strategic attention on word repetitions or fixating only on a fragment of the whole message.

## Background

In recent years, the brewing anxiety over the cognitive validity of listening tests has translated into the emergence of research concerning the correlation between test-takers' listening performance tests scores or similar tasks and their test taking behaviors (Aryadoust, 2019; Field, 2009; Holzknecht, 2017; Suvorov, 2013, 2015; Winke & Lim, 2014). Notably, Field (2009) highlighted two types of listening behaviors deviating from the normal listening process unique to listening under while-listening performance test conditions: strategic behavior and task-specific behavior. The former prepares the listener for the upcoming task "to maximize the amount retained or to compensate for problems of understanding", while the latter involves "attempts to exploit loopholes in the format of the task" to achieve a better score (Field, 2013, p. 27). Aptly, research conducted by Field (2009) on the cognitive validity of listening tests using recorded lectures seems to support this proposition. Field found that while-listening performance test-takers use keywords from the written text as signals to locate information in the question paper prior to listening to the recorded lecture. More interestingly, they listened out for one-to-one matches or paraphrased variations of a written statement in order to select their answers (Field, 2009). Test-takers also declared that the available options in multiple choice questions prompted them to listen out for the "spoken forms, associates or synonyms of key words which they had seen in written form" (Field, 2009, p. 35). This implies that the multiple choice question type probably encourages test-takers to verify information discerned from the listening input in the recorded lecture against pre-established written prompts in the question paper. Field (2009) thus concluded that test-takers dedicated a significant amount of attention to item wording in the question paper and

engaged in the test-wise strategy called keyword-matching, rather than assimilating the overall meaning of the recording. Badger and Yan's (2009, p. 467) findings also vouch for the tactics undertaken by while-listening performance test-takers such as listening out for details and keywords and deducing answers by using information from the text and the examination question paper, thus, again, working out answers by means of "[exploiting] loopholes in the format of the task" (Field, 2013, p. 27). In sum, while such listening assessment instruments are ostensibly a listening test, their construct can be flawed, or at least influenced in unknown ways, due to the dependence on reading ability or test-takers' overreliance on the visual modality, which is a source of inauthenticity in these types of assessments.

Specifically in a while-listening performance test like the CAEL Computer Edition (CE) Online Course, the instructor advises prospective test-takers to listen out for keywords in test items that will help them find the answer while listening to the lecture (CAEL Test, 2018). Such over-reliance on the written text discounts the effectiveness of the tests, since they assess test-taking strategies and depend on reading more than listening skills (see Badger & Yan, 2009). Goh (2002) hypothesized that test-takers who fail to "hear or note down a sufficient number of keywords" (p. 8) will not be able to fully grasp the meaning of the audio message. Furthermore, the keyword-matching strategy could be counter-productive if a listener misses out a keyword match and continues to listen out for that particular word thus missing out subsequent ones (Field, 2009).

**The Present Study**

While Field (2009) asserted that certain words attracted the attention of test-takers who then employed keyword-matching in answering test items, he did not explain

what is meant by keywords. As discussed earlier, there is limited research that investigated the role of keyword-matching as a predictor of while-listening performance test-takers' performance (Aryadoust, 2019; Badger & Yan's, 2009; Field, 2009). Thus, this study aims to investigate how keywords in the test items are attended to by test-takers, and to examine the relationship between such keyword-matching strategies and while-listening performance test performance, by using eye-tracking technology to measure test-takers' gaze behaviors.

Eye tracking has become an increasingly popular means of research in listening assessments to investigate the cognitive mechanisms of while-listening performance test-takers or similar tasks (Aryadoust, 2019; Holzknecht et al., 2020; Suvorov, 2013, 2015; Winke & Lim, 2014). For example, Suvorov (2015) used eye-tracking to examine the relationship between test-takers' viewing behavior in second language listening assessments and test performance. More recently, Holzknecht et al. (2020) also applied eye-tracking technology to investigate the influence of order and spatial location of multiple-choice options on listeners' viewing behavior in a listening test. In another recent study, Batty (2020) found that listeners attended to visual cues to detect the emotional status of the characters in a video-based listening test.

In the present study, gaze behaviors are represented as fixations (dwells) and visits. Fixations are defined as "eye movements that stabilize the retina over a stationary object of interest" (Duchowski, 2007, p. 46) which allows the test-taker to gather information, while visits refer to the time when the test-taker's eye falls on a part of the text until they look away from that part (Aryadoust, 2019). Fixations and visits are measured by duration and counts for all test-takers. Therefore, instead of a

coarse-grain impression of keyword-matching in listening tests, this study examines in detail exactly which words test-takers pay more attention to and how this influences their test performance, i.e., which words helped them score on test items and which distracted them from other ostensibly more consequential keywords.

We further defined the areas of interest (AOIs) in the test items. AOIs are demarcated areas that contain objects of interest to this study (Goldberg, et al. 2002). As this study is concerned with keywords, AOIs are marked out as grammatical classes of words that are potentially deemed as keywords. We defined keywords as all nouns, verbs, and adjectives found within each question and its available multiple-choice question (MCQ) options since they often carry the bulk of meaning (Howell et al., 1999; Reed, 2000). Nouns include both proper nouns and common nouns, which are words that name people, quality, objects, places, ideas, and activities (Oxford Learner's Dictionary, n.d.). Verbs are words that convey actions, states, and processes. Only main lexical verbs were examined while auxiliary verbs such as be-verbs and modal verbs were not considered. Adjectives are descriptive words that attribute qualities to nouns (Oxford Learner's Dictionary, n.d.). These lexical categories were selected according to previous research on gaze behavior with regards to content and function words (Howell et al., 1999; Rayner, 2009; Schotter et al., 2012).

Content words are fixated on about 85% of the time, while function words are only fixated on about 35% of the time (Rayner, 2009; Schotter et al., 2012). Since function words are typically glossed over, the absence of gaze data for these cells would render any conclusions drawn from them somewhat suspect. Moreover, Howell et al. (1999) underscored that content words are critical in communicating meaning

and Angelis (2005) corroborated such findings, asserting that content words "carry more semantic weight than function words" (p. 397). This implies that content words, rather than function words, are considered meaningful parts of speech and are thus devoted more attention (Angelis, 2005). Similarly, Halliday (1985) also measured lexical density only by means of content words like nouns, verbs, and adjectives in his early study. All function words including pronouns, determiners, conjunctions, and prepositions were therefore not included as key words in the present research as, according to previous research, they contain less semantic content and serve mainly grammatical or functional purposes (Angelis, 2005).

Thus, by adopting eye tracking technology, the current study aimed to determine whether the construct of CAEL listening test is significantly affected by reading, which is a construct-irrelevant factor. The research questions (RQs) of the current study are, as follows:

RQ1: How do test-takers attend to keywords in the listening test items of CAEL?

RQ2: How does keyword-matching influence their listening test performance?

The study also aimed to apply a commonly used psychometric measurement model, the Rasch-Andrich model, for psychometric validation. By comparing the results of these two test validation methods, we were able to determine the applicability and drawbacks of each method.

## Method

### Participants

The current study involved 66 students aged between 24 and 40 ($M$ = 25.98, $SD$ = 6.24) and from a major university in Asia. The participants were recruited via posters

and a social media platform (i.e., Telegram). All were healthy, with no reported learning disabilities or neurocognitive disorders. The majority spoke English as their first language (65%). These native speakers were included as a control group to provide baseline performance since they were less likely to be test-wise. Sixty-two percent of the participants were from Singapore (62%), while the remaining 38% comprised a variety of nationalities including Chinese, Indonesian, Burmese, Canadian, Indian, and Vietnamese. All participants provided informed consent and were compensated $10 at the end of the study. This study obtained approval from the academic ethics committee of the University (approval ID: IRB-2020-04-028).

**The computer-mediated listening test**

The participants undertook one section of a retired CAEL CE test (i.e., the Part 3 Long Listening). The listening text was a lecture from an Economics introductory course, and the test comprised 11 test items that were presented across five pages, of which eight were MCQs, two were multiple-choice gap-fill questions, and one was a multi-select question. All MCQs and multiple-choice gap fill questions comprised four options.

We first analyzed the content of the items, wherein we found that there could be local dependence between items 4 and 5. Test-takers who would answer item 5 correctly ("A diagram is a type of ____ model." (graphic)) could be cued to the answer to the previous item ("What is one way of grouping economic models that the professor mentions?" (A: mathematical and visual)). Success on these items seems to be contingent on reading and understanding that "visual" and "graphic" are synonymous. Moreover, items 4 and 5 occur in the same section, so test-takers could have the option to alter item 4 responses immediately after viewing item 5. Therefore,

we set out to investigate whether there were significantly more visits and longer visit durations on AOI at items 4 and 5.

In analyzing the content of the rest of the items, we were looking for evidence that test-takers would respond correctly by simply reading the items in which case the primary relevant construct on CAEL would be reading rather than listening. Overall, we noticed that only 4 of 11 items seemed to actually require listening to a passage to respond correctly while the other items could be answered by reading and applying commonly known introductory economics knowledge. If this speculation is correct, then success on the test would be contingent on non-construct related factors—specifically reading ability and/or visual search strategies. We applied eye tracking to showcase a novel empirical method to verify our content analysis findings.

**Data collection procedures**

The experiment was conducted in a quiet, well-lit room at the university where the study was conducted. The listening test was hosted on a *.html* website via Tobii Pro Studio Version 3.4.8. It was presented on a secondary 23-inch widescreen monitor, which was connected to the primary laptop (HP Pavilion 15, Hewlett Packard, CA, USA) and the Tobii Pro TX300 eye-tracker (Tobii, 2017) that recorded participants' eye-movements binocularly at 300Hz, with 0.6° accuracy. At the start of each session, all participants were seated approximately 65cm from the monitor, with a set of keyboard and mouse in front of them. Subsequently, a nine-point calibration procedure was performed to establish the participants' gaze in relation to the monitor. Following calibration, the test began. To reduce drifts, we advised all participants to minimize large head movements during the tests.

The test consisted of three phases: pre-listening, while-listening, and post-listening. In the pre-listening phase, participants had 2.5 minutes to read the questions. Subsequently in the while-listening phase, participants were required to listen to a mini-lecture (duration = 5.5 minutes). Participants were required to listen to the mini-lecture and fill the electronic answer sheet. In the post-listening phase, participants were given 3 minutes to complete any unanswered questions and/or to amend their answers.

## Data Processing

### *Item scores*

All items were scored according to the answer key that was provided by the CAEL CE test developer – Paragon Testing Enterprises. For the MCQs and multiple-choice fill-in-the-blanks, one mark was given to each correct response and zero mark was given for each incorrect response (including blanks). For the multiselect questions, a maximum of two marks was awarded for correct responses, one mark for partially correct responses, and zero marks for incorrect responses (including blanks).

### *Gaze behavioral measures*

As this study focused on examining how participants gazed at nouns, adjectives, and verbs in the test items while listening to the lecture, the sequences for all scenes viewed by the participants during the while-listening stage were firstly annotated in Tobii Pro Studio Version 3.4.8 (Tobii, 2017). Using this software, we drew three types of areas of interests (AOIs), which were classified as nouns, adjectives, and verbs (see Supplementary Table 1 for further details). Subsequently, the raw eye-tracking data were interpolated with a maximum gap length set at 75ms to replace missing data that arose from tracking issues (Komogortsev et al., 2010; Olsen, 2012),

and smoothed using a moving median filter with a window size of 3 samples to reduce noise (Juhola, 1991; Olsen, 2012). Lastly, fixation eye movements were parsed using a velocity threshold-identification filter (Stuart et al., 2019), with velocity threshold at 30°/s (Olsen & Matos, 2012) and a minimum fixation duration at 100ms (Rayner, 1998).

Six variables of interests were then extracted for every AOI across all items including, (i) average fixation duration, (ii) total fixation duration, (iii) fixation counts, (iv) average visit duration, (v) total visit duration, and (vi) visit counts (see Tobii (2016) for variable definitions). Previous research has provided evidence that these gaze measures are appropriate for research purposes that are similar to the present study such as in word recognition (Cunnings et al., 2017), analysis of multimodal input (Pellicer-Sánchez et al., 2020), and evaluating sentence processing and attention to written words (Issa & Morgan-Short, 2019; Keating & Jegerski, 2015).

In the present study, the number of words across the three classes of content words (i.e., nouns, verbs, and adjectives) differed within each test item. To compare the participants' gaze behaviors across the classes of content words fairly (i.e., relative comparisons), the six variables of interests obtained from all words listed under the same class of content words (e.g., Nouns A, B, and C) were summed and then divided by the number of words in each class (e.g., three words, see statistical analysis). Furthermore, absolute comparisons of participant's gaze behaviors across all classes of content words were also performed. All six variables of interest extracted from all words listed under the same class of content words (e.g., Verbs A, B, and C) were summed (see statistical analysis).

## Data Analysis

### *The Rasch-Andrich Rating Scale Model*

This study utilized the Rasch-Andrich rating scale model analysis of dimensionality and reliability. The model is a psychometric mathematical framework based on the item response theory, which homogeneously measures item difficulty and person ability on an interval scale (Ehrich et al., 2015). By rating a set of representative attributes, the model measures item values and individual test performance on a continuum to compute the latent trait tested for, thereby predicting the behavior of items and persons in a testing situation (Bond & Fox, 2015; Deane et al., 2016). Measurements of fit, unidimensionality, local independence, as well as reliability and separation for item and person were computed using Winsteps, Version 4.5.3 (Linacre, 2020).

### *Fit Statistics*

Infit and outfit statistics, which respectively denote inlier-sensitive quality control statistics and outlier-sensitive quality control statistics, were computed for the test items. Following Bond and Fox (2015), we assumed that the mean square (MnSq) indices that exceed 1.4 are underfitting as they suggest unexpected variance, where high-ability test-takers answered the test item incorrectly but low-ability test-takers answered the same item correctly, while MnSq indices that fall below 0.6 are conversely known as overfitting and similarly fail to conform to the unidimensional requirement of the Rasch measurement (see Bond & Fox, 2015, for further explanation).

### *Reliability*

We computed person and item reliability and separation, too. Person and item reliability values exceeding 0.80 suggest consistency in the test-takers' responses (Bond & Fox, 2015), although reliability values below 0.8 are not indicative of poor test function, but perhaps of small variance in largely homogenous samples of persons and/or items. In addition, we computed separation indices which indicate the number of levels of item difficulty or person ability found by the Rasch-Andrich rating scale model. Particularly, person separation index shows the number of statistically distinct ability groups of test-takers, while item separation index shows the number of statistically distinct item difficulty. Person separation values exceeding 1.00 provides evidence that test-takers are "measured across the continuum" (Krishnan & Idris, 2014, p. 55).

***Unidimensionality and local independence***

We examined psychometric unidimensionality to evaluate whether the test assesses only one specific dimension (Ariffin et al., 2010), which in this study is the while-listening performance construct. Values for raw variance explained by measures above 20% as acceptable, above 40% as good and above 60% as excellent, indicating that the test instrument is psychometrically unidimensional. At the same time, values for unexplained variance should not exceed 15% (Bond & Fox, 2015). In addition, local independence was evaluated by examining the correlation among the residuals of the Rasch-Andrich rating scale model. (Although the preceding content analysis suggested the possibility of local independence, The Rasch-Andrich rating scale model did not provide psychometric evidence to support this assumption).

**Statistical Procedures**

We examined the suitability of two commonly used statistical models to explain the observed variance in the data: linear mixed effect model (LMM) and generalized linear model (GLM). The statistical analysis procedures are discussed in detail next.

***Linear mixed effect model***

We first used the `lme4` package (Bates et al., 2020) in *`RStudio`*, Version 2022.02.3, to conduct a series of linear mixed effect models (LMMs) to examine whether there was any evidence for data clustering within the test items and persons (Brown, 2021). A visual inspection of the data suggested that there might be some clustering (grouping) present in the data, which would warrant further examination. That is, the predictive effect of the gaze measures in the study vary according to test items when clustering is present.

As Figure 1 demonstrates, there were some (overlapping) clusters in the test items' total scores plotted against the average total fixation duration (TotFixDurAvg) on nouns, adjectives, and verbs (the figures in the left column). The regression lines in each graph possess different slopes and intercepts, although they do not seem to create conspicuous groups in the data. Similar patterns were observed when the rest of the gaze measures mentioned above were plotted, thereby warranting an LMM analysis in the study.

In this study, LMMs consisted of both fixed and random effects. Fixed effects are the independent variables that are constant across individuals and, in our research, consist of the 12 gaze measures that were drawn from the eye tracking study (Table

1). Random effects, on the other hand, vary across groups/clusters of individuals and, in the present study, comprise test items (Table 1).

As a result, we tested two LMMs, which are demonstrated in Table 1. While the models consist of 12 fixed effects, they are different in terms of their random effects. The random effects for the models comprise of items as a random intercept (Model 1) and as a random intercept and a random slope (Model 2), respectively. This indicates that the intercept of Model 1 varies by the grouping effect or test items; in the same vein, the intercept and slope of Model 2 vary by the grouping effect or test items. Therefore, instead of having one regression line for all the items and persons in these models, we generated 11 regression lines each representing one listening test item with varying intercepts. In Model 1, only the intercept of the regression lines varied, while in Model 2, both the intercept and slope of the models varied (see Brown, 2021, p. 10, for details).

**Table 1**

*Fixed and random effects of the four LMMs*

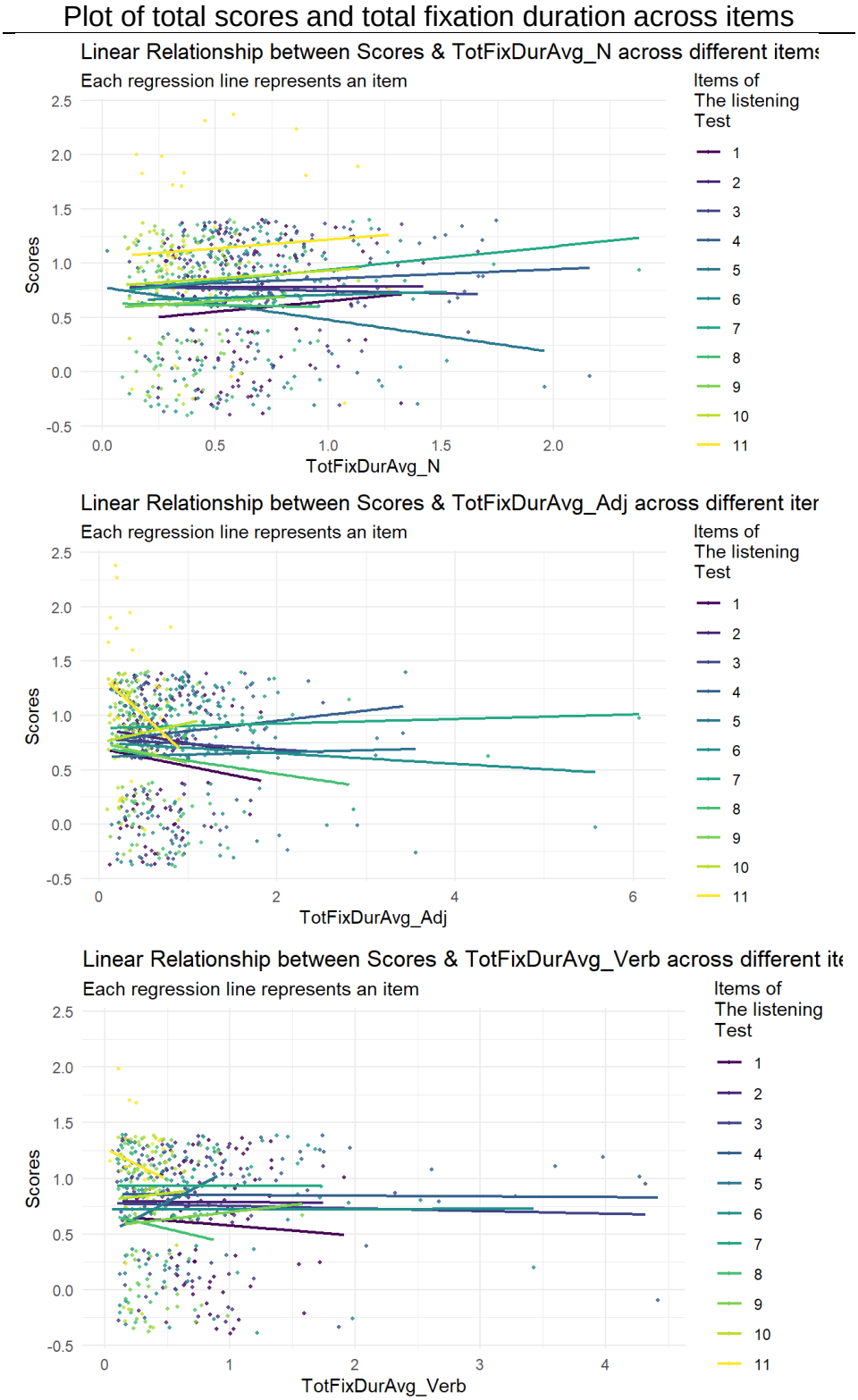
| Variable | Model 1 | Model 2 |
|---|---|---|
| TotFixDurAvg_N | Fixed | Fixed |
| FixCountAvg_N | Fixed | Fixed |
| TotVisDurAvg_N | Fixed | Fixed |
| VisCountAvg_N | Fixed | Fixed |
| TotFixDurAvg_Adj | Fixed | Fixed |
| FixCountAvg_Adj | Fixed | Fixed |
| TotVisDurAvg_Adj | Fixed | Fixed |
| VisCountAvg_Adj | Fixed | Fixed |
| TotFixDurAvg_V | Fixed | Fixed |
| FixCountAvg_V | Fixed | Fixed |
| TotVisDurAvg_V | Fixed | Fixed |
| VisCountAvg_V | Fixed | Fixed |
| Items | Random intercept | Random intercept & slope |
| Persons | | |
| Scores | Dependent variable | Dependent variable |

*Note*: Adj = adjective; Avg = average; Dur = duration; Fix = fixation; N = noun; Tot = total; V = verb.

**Figure 1**

*Plot of Total Scores and Total Fixation Duration across Items with Separate Regression Lines*

(Legend: Adj = adjective; Avg = average; Dur = duration; Fix = fixation; N = noun; Tot = total)



**Generalized linear model (GLM) binary probit regression**

*Procedures.* Next, respective regression models (i.e., binary probit regression analysis on test items 1 to 10 and multinomial logistic regression on test item 11) were generated, adopting four ways for every test item using the following independent variables:

Procedure 1: Collective average of eye-tracking variables, i.e., the average of all fixation variables was processed together as co-variates in a single analysis. The same procedure was repeated on Visit variables.

Procedure 2: Individual averages of eye-tracking variable, i.e., the averages of each individual fixation variable (average fixation duration, total fixation duration and fixation count), were processed one by one as distinct co-variates in separate analyses. The same procedure was repeated on Visit variables.

Procedure 3: Collective content words, e.g., Fixation Duration for all content words belonging to an AOI (e.g., Verbs A, B, C and D) were processed together as co-variates in a single analysis, followed by Total Fixation Duration and Fixation Count. The same procedure was repeated on Visit variables.

Procedure 4: Individual content words, e.g., Fixation Duration for each content word (e.g., Noun A, Noun B, Noun C and Noun D) were processed one by one as distinct co-variates in separate analyses, followed by Total Fixation Duration and Fixation Count. The same procedure was repeated on Visit variables.

For the two test items in the format of fill-in-the-blank questions with drop down menus (Q5 and Q8), care was taken to separate scenes in which the drop-down menus were opened (OP), i.e., when test-takers were able to view the MCQ options and choose their answers, and closed (CL), i.e., when test-takers were unable to view the MCQ options. In scenes where the drop-down menus were closed, we also considered whether it was before or after test-takers had chosen their answers. This entails that scenes with empty blanks were processed separately from those with blanks filled in by the test-taker's chosen MCQ option. The chosen answer also influenced our consideration of scenes for succeeding test items on the same page. This means that if a test-taker had chosen option A for Q5, the scenes for Q6 and Q7 will follow the scene named after option A, e.g., FixDur_CL_Ans_graphic_Q6_A1_noun_factor, which describes the Fixation Duration data for the noun "factor" in the first MCQ option of Q6, where the drop-down box for Q5 is closed and the test-taker had answered "graphic". The findings show that test-takers' performance in all 11 items can be attributed to gaze measures.

*Binary probit regression.* We used IBM SPSS for Windows Version 20 (IBM Corporation, 2011) to carry out generalized linear models connect linear models. We applied a link function to make the variance of each measurement a function of its predicted value. These models allow for response variables with arbitrary distributions (i.e., non-normal distributions) and arbitrary functions of the response variable (i.e., the link function) to vary linearly with the predicted values. Specifically, the binary probit regression is an SPSS annotated output developed to process such data deviating from the standard normal distribution, "where a binomial variate has a parameter related to an assumed underlying tolerance distribution".

Binary probit regression measures the relationship between the influence of a stimulus and the ratio of cases demonstrating a particular reaction to the stimulus (Field, 2018) – the stimulus in this case being the fixation and visit data per test-taker on each class of content words per test item and the response being the test score per test-taker per test item. This procedure is especially appropriate for designed experiments using experimental data, to test if the dichotomous output is influenced by the independent variables (Field, 2018). Thus, we utilized binary probit regression to analyze data for 10 out of 11 test items which produce binary outputs. Using the inverse of the cumulative standard normal distribution function, the procedure reports measures of effective values for various rates of response (Field, 2018). We estimated an omnibus chi-squared test that shows whether there is a significant difference in independent variables across test-takers' who answered test items correctly and those who answered incorrectly ($p < .05$).

*Multinomial logistic regression.* As the multi-select question (i.e., test item 11) was scored using partial credit, a multinomial logistic regression was used as item scores could fall into one of three categories (i.e., incorrect, partially correct, or correct response). Given a covariate pattern, responses are assumed to be independent multinomial variables (Field, 2018).

## Results

### Rasch-Andrich Rating Scale Model Analysis

The Rasch-Andrich rating scale model analysis results showed the infit and outfit MnSq indices for person to be 0.96 and 1.21 respectively, while that for item to be 0.97 and 1.21, thus suggesting a good fit.

The analysis results for this study found the person reliability and person separation index to be 0.57 and 1.15 respectively, indicating a hierarchy of ability across test-takers despite their homogeneity, redeemed by the fact that test items were targeted at the ability level of these test-takers (Bond & Fox, 2015). Meanwhile, Bond and Fox (2007) suggest that item reliability values exceeding 0.80 are strongly acceptable. Krishnan and Idris (2014) stated that an item separation value exceeding 1.00 indicates that "items have enough breadth" (p. 55), while Linacre (2003) stated that item separation values exceeding 2.00 are considered good. The Rasch-Andrich rating scale model analysis results for this study found item reliability and item separation index to be 0.98 and 6.95 respectively, indicating that the operationalized construct is excellently represented and caters to very diverse ability levels.

Through the principal component analysis of residuals (PCAR), the Rasch-Andrich rating scale model analysis found an excellent raw variance value for this study at 72.3 Eigenvalues, or 76.7%, thus implying that the listening test analyzed indeed measured the single trait. Considering the fit indices, reliability and separation values and PCAR results for unidimensionality, the analysis provides evidence that the test is psychometrically "sound." Nevertheless, this does not nullify the possibility that test-takers deployed construct-irrelevant processes such as keyword matching to answer the test items, as psychometric dimensions (created by methods such as Rasch measurement in general) are not necessarily compatible with the psychological reality of test taking.

**Linear Mixed Effect Models (LMMs)**

While the LMMs converged, they did not show any significant effects of the fixed and random variables on the test scores. Table 2 demonstrates the fit statistics and

predictive power of the models. As attested by the AIC and BIC fit statistics, Model 1 has a better fit to the data (as indicated by the small AIC and BIC indices). Nevertheless, the interclass correlation (ICC), which is the amount of variance explained by clustering structures, was negligible in both models. In addition, the $R^2$ marginal and conditional indices were very low and negligible in these models. Overall, the LMM results indicated that neither of the models would be suitable for describing the observed variance in the data. Thus, a series of GLM analyses was conducted to determine what gaze measure variables could explain variance in the test scores.

Table 2
*Fit Statistics and Predictive Power of the Models*

| Fit index | AIC | BIC | $R^2$ marginal | $R^2$ conditional | Convergence | Variance of random effects | ICC |
|---|---|---|---|---|---|---|---|
| Model 1 | 557.6302 | 623.3781 | 0.0113 | 0.0453 | Yes | 0.00658 | 0.0345 |
| Model 2 | 561.8707 | 627.6186 | 0.011 | 0.0287 | Yes | 0.00342 | 0.018 |

Note: * Due to space constraints, only the variance of the intercept is displayed for this model.
** Due to the large number of parameters, Model 4 did not converge.

**Probit and Multinomial Regressions**

For GLM procedure 1, test-takers' performances in 2 out of 11 items could be attributed to the collective average of eye-tracking variables (see Table 3). The binary probit regression omnibus test found a significant difference ($p < .05$) for average visit variables on nouns in test item 7 while average fixation and visit variables for adjectives in test item 4 and average fixation variables for nouns in test item 7 were shown to be approaching significance. The multinomial logistic regression did not show significant findings for test item 11.

Table 3

*Results from Procedure 1 – Collective Average of Eye-Tracking Variables*

| **Test items** | **Classes of content words** | **Eye-tracking variables** | **Omnibus significance (*p* value)** | **Correct responses M (SD)** | **Incorrect responses M (SD)** | **Number of included cases *N* (%)** |
|---|---|---|---|---|---|---|
| 4 | Adjective | Average fixation variables | .054* | 7.77 (3.61) | 7.31 (5.61) | 63 (96%) |
| | | Average visit variables | .056* | 11.10 (5.15) | 10.40 (7.97) | |
| 7 | Noun | Average fixation variables | .057* | 5.90 (2.24) | 6.42 (2.93) | 64 (97%) |
| | | Average visit variables | .011 | 7.40 (2.95) | 7.73 (3.43) | |

*Note.* **p* values approaching significance.

As demonstrated in Table 4, for procedure 2, test-takers performances in 3 out of 11 items (27%) were attributed to the individual average of eye-tracking variables. The binary probit regression omnibus test found a significant difference ($p < .05$) for average fixation and visit count for adjectives in test item 4, average fixation and visit duration for verbs in test item 5, average total fixation duration, fixation and visit count for nouns in test item 7, while the average fixation count for nouns in test item 7 was shown to be approaching significance. The multinomial logistic regression test found a significant difference ($p < .05$) for visit duration, total visit duration and visit count on nouns in test item 11 values, while the average fixation duration, total fixation duration and fixation count were shown to be approaching significance.

Table 4

*Results from Procedure 2 – Individual average of eye-tracking variable*

| **Test Items** | **Classes of content words** | **Eye-tracking variable** | **Omnibus significance (*p* value)** | **Incorrect responses M (SD)** | **Partially correct responses M (SD)** | **Correct responses M (SD)** | **Number of included cases *N* (%)** |
|---|---|---|---|---|---|---|---|
| 4 | Adjective | Average Fixation Count | .038 | 2.34 (1.460) | - | 3.57 (2.170) | 63 (96%) |
| | | Average Visit Count | .033 | 2.22 (1.420) | - | 3.41 (2.060) | |
| 5 | Verb | Average Fixation Duration | .044 | 0.190 (0.065) | - | 0.292 (0.125) | 19 (29%) |
| | | Average Visit Duration | .044 | 0.190 (0.065) | - | 0.292 (0.125) | |
| 7 | Noun | Average Total Fixation Duration | .016 | 0.429 (0.231) | - | 0.706 (0.388) | 64 (97%) |
| | | Average Fixation Count | .009 | 1.82 (0.943) | - | 2.87 (1.140) | |
| | | Average Visit Duration | .055* | 0.412 (0.450) | - | 0.253 (0.063) | |
| | | Average Visit Count | .013 | 1.74 (0.975) | - | 2.65 (0.981) | |
| 11 | Noun | Average Fixation Duration | 0.100 | 0.236 (0.096) | 0.242 (0.045) | 0.242 (0.058) | 39 (59%) |
| | | Total Average Fixation Duration | 0.070 | 0.473 (0.360) | 0.414 (0.223) | 0.503 (0.326) | |
| | | Fixation Counts | 0.090 | 1.90 (1.210) | 1.77 (0.962) | 2.05 (1.290) | |
| | | Average Visit Duration | 0.012 | 0.251 (0.111) | 0.261 (0.068) | 0.245 (0.058) | |
| | | Total Average Visit Duration | 0.009 | 0.482 (0.376) | 0.418 (0.223) | 0.504 (0.327) | |
| | | Visit Count | 0.011 | 1.73 (0.991) | 1.65 (0.887) | 1.99 (1.220) | |

*Note.* * p values approaching significance

For procedure 3, test-takers performances in 3 out of 11 items were attributed to gaze measures on collective content words (see Table 5). The binary probit regression omnibus test showed a significant difference ($p < .05$) for fixation duration on verbs in test item 1, fixation and visit count for verbs in test item 2, total fixation and visit duration for adjectives in test item 3, while fixation count for adjectives in test item 3 was shown to be approaching significance. The multinomial logistic regression did not show significant findings for test item 11.

Table 5

*Results from Procedure 3 – Collective Content Words*

| **Test Item** | **Classes of Content Words** | **Eye-tracking Variable** | **Omnibus significance (*p* value)** | **Incorrect Responses M (SD)** | **Correct Responses M (SD)** | **Number of included cases (*N*)** |
|---|---|---|---|---|---|---|
| 1 | Verbs | Fixation Duration | .026 | 1.06 (2.890) | 0.973 (0.416) | 30 (46%) |
| 2 | Verbs | Fixation Count | .003 | 8.21 (3.620) | 8.19 (4.110) | 18 (27%) |
| | | Visit Count | .043 | 7.71 (3.270) | 7.62 (3.860) | |
| 3 | Adjectives | Total Fixation Duration | .010 | 17.60 (13.900) | 17.10 (14.300) | 23 (35%) |
| | | Fixation Count | .057* | 4.56 (3.820) | 4.23 (3.660) | |
| | | Total Visit Duration | .009 | 4.63 (3.860) | 4.27 (3.690) | |

*Note.* * *p* values approaching significance

For procedure 4, test-takers performance in all 11 items could be attributed to gaze measures on individual content words. The binary probit regression omnibus

found a significant difference ($p < .05$) for the eye-tracking variables in eight test items on nouns and adjectives respectively, and 3 test items for verbs. Only statistically significant values with 20 cases (30%) or more are reported in Supplementary Table 2, while significant values with fewer than 20 cases and values approaching statistical significance are attached in Supplementary Table 3. The multinomial logistic regression found a significant difference ($p < .05$) for eye-tracking variables on nouns in test item 11, but not on verbs and adjectives.

Binary probit regression analysis showed significant difference for all eye-tracking variables on nouns, verbs, and adjectives. Gaze measures on nouns were the most significant predictors of test-takers' performance, predicting scores in 8.67 (≈ 9) out of 11 test items (79%) on average across all eye-tracking variables. Gaze measures on adjectives were the next most significant predictors of test-takers' performance, predicting scores in 5.67 (≈ 6) out of 11 test items (52%) on average across all eye-tracking variables. Gaze measures on verbs, however, were not as significant in predicting test-takers' performance, with only 2 out of 11 test items (18%) on average across all eye-tracking variables.

Finally, as demonstrated in Supplementary Tables 2 and 3, the highest number of significant predictors (19) of test-takers' performance was found in test item 3, with 12 significant values (10 for nouns, 2 for adjectives) and 7 values approaching significance (5 for nouns, 2 for adjectives) distributed among 6 eye tracking variables across 4 nouns and 4 eye tracking variables on an adjective. This was followed closely by test item 4 with 18 significant predictors, with 14 significant values (7 for nouns, 7 for adjectives) and 4 values approaching significance (2 for nouns, 2 for adjectives) distributed among 5 eye tracking variables across 3 nouns and 6 eye

tracking variables on 4 adjectives. Test item 1 and 6 both comprised 11 significant predictors, where the former consisted of 6 significant values (5 for nouns, 1 for verb) and 5 values approaching significance (2 for nouns, 1 for verbs, 2 for adjectives), while the latter consisted 5 significant values (2 for verbs, 3 for adjectives) and 5 values approaching significance (1 for noun, 4 for adjectives). The only test item without any significant predictors is test item 10, where none of the eye tracking variables showed significant difference ($p < .05$) on the AOIs. Overall, the analysis for procedure 4 found that 17 out of 105 nouns (16%), 3 out of 31 verbs (10%) and 15 out of 59 adjectives (25%) across all test items showed a significant difference and accounted for test-takers' performance.

## Discussion

This study sets out to investigate how keywords in the test items are attended to by test-takers, and how keyword-matching influences while-listening performance. Although the Rasch-Andrich rating scale model provided supporting evidence for the psychometric validity of the test, we found evidence that gaze measures on content words, i.e., nouns, verbs, and adjectives, can indeed predict test-takers' performance on this type of listening comprehension test items. In language assessment, the difference between psychometric and psychological dimensions is well-known (e.g., McNamara, 1991). Our results are important as they show that psychometric dimensions like those that are created by Rasch measurement are broad and might consist of psychological (rather than psychometric) sources of construct-irrelevant variance which cannot be captured by psychometric analysis.

In the present study, total fixation duration and visit duration on nouns were found to be the most significant influencers on test-takers' performance with nine

significant predictors each. This suggests that the longer test-takers' visited and fixated on particular nouns in a test item, the more likely they were to answer the test item correctly. Thus, listening out for nouns during the while-listening performance test had likely played a significant role in focusing test-takers' attention on important parts of the lecture, helping them to locate the answer in the listening input. Consistent with the study conducted by Badger and Yan (2009), it may be said that test-takers engaged in test-specific strategies like using written information from the test items to deduce the answer. Given the findings from our research, this study is the first to provide quantitative evidence for what Badger and Yan (2009) claimed about while-listening performance tests.

In addition, whereas previous research by Field (2009) broadly speculated that keywords played a role in the answering of test items, Field neglected to crystallize the definition of keywords. In the present study, however, we defined keywords as content words consisting of nouns, verbs, and adjectives. As earlier noted, a key finding is that test-takers who answered a test item correctly had paid more attention to nouns found in the test items, followed by adjectives, then verbs. This implies that nouns may be considered the most significant keywords among all content words in test items on while-listening performance tests, while adjectives are fairly significant keywords and verbs only remotely significant. Based on previous research by Angelis (2005) and Halliday (1985), we propose that nouns are easier to process, especially if they are imageable (see Authors, XXXX). By focusing on imageable nouns, listeners can allocate more memory capacity to the overall auditory texts. On the other hand, adjectives and verbs seem to be less imageable and, compared with nouns, more likely to tax working memory while test-takers engage in the sort of multitasking that

while-listening performance tests demand. We call for future research to build on our findings and investigate what makes nouns in these tests have such a significant role in answering test items. As such, the imageability of nouns and its relationship with test difficulty and test-takers' performance should be problematized (i.e., be treated as a research question warranting answers). The implication of this new line of research would be significant for listening assessment. It would shed light on the cognitive processes that may yield 'construct-irrelevant variance' in test scores (Messick, 1996) by examining the interplay between scores (end product of tests) and gaze behaviors of test-takers (test-taking process). In addition, item writers, listening test developers, and practitioners would benefit from such a line of research. If we manage to discover reading-specific processes that influence listening test performance (an enigma hitherto; see Buck, 2001), we will be able to design test items in such a way that the effect of test-takers' reliance on reading could be controlled for. Overall, given that our findings elucidate the pattern and nature of content words which test-takers' pay more attention to, the present study further streamlines the definitions of keywords in the study that was previously reviewed.

Furthermore, our findings have shed light on the intricacies of keyword-matching strategy employed in while-listening performance tests and its influence on test-takers' performance, perhaps harking back to the inadequacy in "cognitive validity" of such listening assessments. This also resonates with a prominent concept in language assessment known as authenticity or cognitive validity (Weir, 2005; Wood, 1993), which refers to how much a test impels language learners to do things (e.g., cognitive processes) which they would not otherwise do in real-life situations. As demonstrated by the results, while-listening performance tests impel test-takers to

exploit loopholes in the task format by engaging in keyword-matching with nouns in test items, which is otherwise irrelevant in real-life listening comprehension. Owing to the test item format, test-takers are influenced to pay attention to how test items are worded so as to check the information perceived from the listening input against written cues in the question paper. As such, keyword-matching seems to threaten the authenticity of listening tests since test scores reflect, inter alia, ability in this test-wise strategy instead of its supposed aim of testing competencies and cognitive processes in listening (e.g., bottom-up and top-down), including both bottom-up and top-down processing (Aryadoust, 2020; Kintsch, 1998). As highlighted, the question of authenticity is critically underresearched in listening assessment and learning (Buck, 2001; Ockey & Wagner, 2018), and significantly more research would be required to address it properly (Douglas, 2001; Ryan & Granville, 2020).

As earlier noted, underlining this threat to while-listening performance test authenticity is the concept of construct-irrelevant variance (Messick, 1994, 1996), which, among other things, implies that test-takers are inhibited from freely demonstrating their listening skills because of restrictively structured test items or response formats. This occurs when the assessment contains "excess reliable variance that is irrelevant to the interpreted construct" (Messick, 1994, p. 8), such that variance in test-takers' performance is not only attributed to person ability under assessment, but also another construct which confounded the assessment, thus contaminating test scores (Messick, 1996). Given that keyword-matching was found to have played a statistically significant role in test-takers' performance, despite being an unintended predictor of test scores, it might be said it constitutes a source inducing construct-irrelevant variance. If the validity of such listening tests is compromised due to

construct-irrelevant variance, then any demonstration of language competencies associated with test performance is at best circumstantial. This has consequential, and possibly even dangerous, implications for language assessments, considering that some of these listening tests bear a significant weight in high-stakes testing in major nationwide and international language examinations. They, for example, affect test-takers' future trajectory as they either allow or deny access to future academic or career opportunities (see Alderson & Wall, 1993, for an argument). If while-listening performance tests fail to discriminate between listening comprehension skills and strategies like keyword-matching, it may misconstrue the listening construct and how test scores are interpreted by stakeholders, therefore likely reducing the validity of such listening assessments.

An upshot of this study is that listening comprehension assessments may avoid the test formats that encourage test-takers to engage in test-specific strategies such as keyword-matching, since listening test-takers' performance should not be influenced by irrelevant measures that interfere with their demonstration of language competence. Future research on listening assessments should also take into consideration the effect of such variables on test-takers' performance. Consequentially, this study brings greater illumination onto the concept of construct and validity in listening assessments which may help in interpreting test scores. As there are no other similar fine-grained investigations carried out on keyword-matching on while-listening performance tests, this study has addressed the research gap and it is recommended that future research expand on this. In addition to recruiting adult university students and staff, this study can be replicated on primary or secondary school-going students to investigate the prevalence of keyword-matching as a test-

specific strategy amongst adolescents currently experiencing the rigour of formal education. Future studies can also look into how examination conditions shape test-taking behavior, to offer a more exact representation of listening comprehension assessments in high-stakes testing.

## Conclusion

This study investigated how keywords may be defined, how keywords in test items are attended to by test-takers and how keyword-matching influences performance on while-listening performance tests. Given our hypothesis that keyword-matching confounds test performance, we specified keywords as content words consisting of nouns, verbs, and adjectives, and used eye-tracking technology to capture the gaze behaviour of 66 test-takers, in terms of fixation and visit variables. Our results showed that keyword-matching is indeed a significant predictor of test-takers' performance and that nouns were the most significant amongst all the content words in influencing test scores. As such, keyword-matching indeed helps test-takers achieve higher scores, hence confirming our hypothesis that keyword-matching is a correlate of test-takers' performance in while-listening performance tests. These findings also provide support for the overwhelming presence of construct-irrelevant variance (which was not captured by the Rasch-Andrich rating scale model), challenging the cognitive validity of while-listening performance assessments.

Based on the findings, it might be said that a major portion of the construct in this CAEL while-listening performance test is reading. Test developers should apply techniques that are sensitive to test taking processes, such as eye tracking, and not limit their validation efforts to the psychometric analysis of the test items. Additionally, test item review by a panel of outside neutral experts must become a

standard procedure to prevent this type of non-construct related influence. Finally, we call for further research into keyword-matching and particularly the differences between the effect of gazing at and fixating on nouns as opposed to adjectives and verbs on the validity of the interpretations and uses of test scores. Understanding how these mechanisms interact with test-takers' cognitive processes in listening assessments will provide evidence of authenticity and cognitive validity. Notably, the present study adopted fixation / visit counts and duration. We suggest that future researchers use first fixation duration and rereading duration to examine the different stages of reading process during listening. It is hoped that the findings of this study will be extended to contexts beyond while-listening performance tests, specifically to the environments wherein listening under non-assessment conditions has an essential role in learning and interaction such as lecture comprehension at universities and other academic contexts.

**Data availability**

The datasets generated during and/or analyzed during the current study are not publicly available as they are the property of [masked] University.

**Abbreviations**

Adj = adjective
AOI = areas of interest
Avg = average
CAEL = Canadian Academic English Assessment
CE = CAEL Computer Edition
Dur = duration
Fix = fixation
GLM = generalized linear model
IELTS = International English Language Testing System
LMM = linear mixed effect model

MCQ = multiple-choice question

MnSq = mean square

N = noun

PCAR = principal component analysis of Rasch residuals

Tot = total

## References


Angelis, G. D. (2005). Interlanguage transfer of function words. *Language Learning, 55*(3), 379-414. https://doi.org/10.1111/j.0023-8333.2005.00310.x

Ariffin, S. R., Omara, B., Isaa, A., & Sharif, S. (2010). Validity and reliability multiple intelligent item using Rasch measurement model. *Procedia Social and Behavioral Sciences, 9*, 729-733. https://doi.org/10.1016/j.sbspro.2010.12.225

Aryadoust, V. (2011). Application of the fusion model to while-listening performance tests. Shiken: JALT Testing & Evaluation SIG Newsletter, 15, 2–9. http://hosted.jalt.org/test/PDF/Aryadoust2.pdf

Aryadoust, V. (2018). The listening test of the internet-based test of english as a foreign language (TOEFL iBT). In D. L. Worthington & G. D. Bodle (Eds.), The Sourcebook of Listening Research: Methodology and Measures (pp. 592–598). Wiley Blackwell.

Aryadoust, V. (2019). Dynamics of item reading and answer changing in two hearings in a computerized while-listening performance test: An eye-tracking study. Computer Assisted Language Learning. https://doi.org/10.1080/09588221.2019.1574267

Aryadoust, V. (2020). A review of comprehension subskills: A scientometrics perspective. System, 88, 102180. https://doi.org/10.1016/j.system.2019.102180

Badger, R., & Yan, X. (2009). The use of tactics and strategies by Chinese students in the Listening component of IELTS. In P. Thompson (Ed.), *International English Language Testing System (IELTS) Research Reports* (Vol. 9, pp. 67-98). British Council and IELTS Australia. https://www.ielts.org/-/media/research-reports/ielts_rr_volume09_report2.ashx

Batty, A. O. (2020). An eye-tracking study of attention to visual cues in L2 listening tests. *Language Testing.* Advance online publication. https://doi.org/10.1177/0265532220951504

Bodie, G. D., Winter, J., Dupuis, D., & Tompkins, T. (2020). The echo listening profile: Initial validity evidence for a measure of four listening habits. *International Journal of Listening, 34*(3), 131-155, DOI: 10.1080/10904018.2019.1611433

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Brown, V. A. (2021). An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science, 4*(1), Article 2515245920960351. https://doi.org/10.1177/2515245920960351

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Cunnings, I., Fotiadou, G., & Tsimpli, I. (2017). Anaphora resolution and reanalysis during L2 sentence processing: Evidence from the visual world paradigm.

*Studies in Second Language Acquisition*, *39*(4), 621–652. https://doi.org/10.1017/S0272263116000292

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*(4), 422-433. https://doi.org/10.3758/BF03214546

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2016). Development of the Statistical Reasoning in Biology Concept Inventory (SRBCI). *CBE Life Sciences Education*, *15*(1), ar5-ar5. https://doi.org/10.1187/cbe.15-06-0131

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing*, *18*(2), 171-185. https://doi/abs/10.1177/026553220101800204?journalCode=ltja

Duchowski, A. (2007). Taxonomy and Models of Eye Movements. In *Eye Tracking Methodology* (pp. 41-48). Springer.

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal, 77*(2), 180–191. https://doi.org/10.2307/328942

Ehrich, J. F., Howard, S. J., Tognolini, J. S., & Bokosmaty, S. (2015). Measuring attitudes toward plagiarism: issues and psychometric solutions. *Journal of Applied Research in Higher Education, 7*(2), 243-257. https://doi.org/10.1108/JARHE-02-2014-0013

Field, A. P. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.

Field, J. (2009). The cognitive validity of the lecture-based question in the IELTS listening paper. In P. Thompson (Ed.), *International English Language Testing System (IELTS) Research Reports 2009* (Vol. 9, pp. 17-65). British Council and IELTS Australia.

Field, J. (2013). Cognitive validity. In A. Garanpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening*. Cambridge University Press.

Geranpayeh, A., & Taylor, L. (2008). Examining listening: Developments and issues in assessing second language listening. Cambridge ESOL: *Research Notes, 32,* 2-5.

Goh, C. C. M. (2002, 2002/06/01/). Exploring listening comprehension tactics and their interaction patterns. *System, 30*(2), 185-206. https://doi.org/https://doi.org/10.1016/S0346-251X(02)00004-0

Haarmann, H. J., Davelaar, E. J., & Usher, M. (2003, 2003/02/01/). Individual differences in semantic short-term memory capacity and reading comprehension. *Journal of Memory and Language, 48*(2), 320-345. https://doi.org/https://doi.org/10.1016/S0749-596X(02)00506-5

Halliday, M. A. K. (1985). *Spoken and written language*. Oxford University Press.

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., & Dunlea, J. (2020). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing*. Advance online publication. https://doi.org/10.1177/0265532220917316

Howell, P., Au-Yeung, J., & Sackin, S. (1999). Exchange of stuttering From function words to content words with age. *Journal of Speech, Language, and Hearing Research, 42*(2), 345-354. https://doi.org/10.1044/jslhr.4202.345

IBM Corporation. (2011). *IBM SPSS Statistics for Windows.* (Version 20) [Computer software]. IBM.

Issa, B. I., & Morgan-Short, K. (2019). Effects of external and internal attentional manipulations on second language grammar development: An eye-tracking study. *Studies in Second Language Acquisition*, *41*(2), 389–417. https://doi.org/10.1017/S027226311800013X

Keating, G. D., & Jegerski, J. (2015). Experimental designs in sentence processing research: A methodological review and user's guide. *Studies in Second Language Acquisition*, *37*(1), 1–32. https://doi.org/10.1017/S0272263114000187

Juhola, M. (1991). Median filtering is appropriate to signals of saccadic eye movements. *Computers in Biology and Medicine*, *21*(1-2), 43-49. https://doi.org/10.1016/0010-4825(91)90034-7

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., & Gowda, S. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering, 57*(11), 2635-2645. https://doi.org/10.1109/TBME.2010.2057429

Krishnan, S. & Idris, N. (2014). Investigating reliability and validity for the construct of inferential statistics. I*nternational Journal of Learning, Teaching and*

*Educational Research, 4*(1), 51-60. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.677.9940&rep=rep1&type=pdf

Linacre, J. M. (2020). *Winsteps.* In (Version 4.5.3) [Computer software]. Winsteps.com

Maftoon, P., & Alamdari, E. F. (2020) Exploring the effect of metacognitive strategy instruction on metacognitive awareness and listening performance through a process-based approach. *International Journal of Listening, 34*(1), 1-20. DOI: 10.1080/10904018.2016.1250632

Masrai, A. (2020). Can L2 phonological vocabulary knowledge and listening comprehension be developed through extensive movie viewing? The case of Arab EFL learners. *International Journal of Listening, 34*(1), 54-69. DOI: 10.1080/10904018.2019.1582346

Messick, S. (1994). A*lternative modes of assessment, uniform standard of validity.* ETS Research Reports. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1994.tb01634.x

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256. https://doi.org/10.1177/026553229601300302

Nation, I. S. P., & Newton, J. (2008). *Teaching ESL/EFL listening and speaking*. Routledge.

Ockey, G. J., & Wagner, E. (2018). *Assessment of L2 listening: Moving towards authenticity*. John Benjamins Publishing Company.

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

Olsen, A., & Matos, R. (2012). *Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies.* ETRA '12: Proceedings of the Symposium on Eye Tracking Research and Applications, Santa Barbara, CA, USA.

Oxford Languages and Google - English. (n.d.). https://languages.oup.com/google-dictionary-en/

Pellicer-Sánchez, A., Tragant, E., Conklin, K., Rodgers, M., Serrano, R., & Llanes, Á. (2020). Young learners' processing of multimodal input and its impact on reading comprehension: An eye-tracking study. *Studies in Second Language Acquisition, 42*(3), 577–598. https://doi.org/10.1017/S0272263120000091

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bullentin, 124*(3), 372-422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search, *The Quarterly Journal of Experimental Psychology*, *62*, 8, 1457-1506, https://10.1080/17470210902816461

Reed, J. (2000). *Assessing vocabulary*. Klett Sprachen GmbH.

Ryan, J., & Granville, S. (2020). The suitability of film for modelling the pragmatics of interaction: Exploring authenticity. *System, 89*, Article 102186. https://doi.org/10.1016/j.system.2019.102186

Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Atten Percept Psychophys*, *74*(1), 5-35. https://doi.org/10.3758/s13414-011-0219-2. PMID: 22042596.

Stuart, S., Hickey, A., Vitorio, R., Welman, S., Foo, S., Keen, S., & Godfrey, A. (2019). Eye-tracker algorithms to detect saccades during static and dynamic tasks: a structured review. *Physiological measurement, 40*(2), 02TR01. https://doi.org/10.1088/1361-6579/ab02ab

Suvorov, R. (2013). *Interacting with visuals in L2 listening tests: An eye-tracking study*. Doctoral dissertation submitted to Iowa State University. https://lib.dr.iastate.edu/etd/13299

Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing*, *32*(4), 463-483. https://doi.org/10.1177/0265532214562099

Tobii AB. (2016). *Tobii Studio user's manual version 3.4.5*. https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-studio-user-manual.pdf

Tobii AB. (2017). *Tobii Pro Studio.* In (Version 3.4.8) [Computer software].

Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics, 19*(4), 432-451. https://doi.org/10.1093/applin/19.4.432

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing, 27*(4), 493–513. https://doi.org/10.1177/0265532209355668

Weir, C. (2005). *Language testing and validation: An evidence-based approach.* Palgrave Macmillan.

Wickens, C. D. (2006). Attention to attention and its applications: A concluding view. In A. F. Kramer, D. A. Wiegmann, & A. Kirlik (Eds.), *Attention: From Theory to Practice* ((pp. 239–249). Oxford Scholarship.

Wilson, M. (2003). Discovery listening—improving perceptual processing. *ELT Journal, 57*(4), 335-343. https://doi.org/10.1093/elt/57.4.335

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, Issue. British Council, Cambridge English Language Assessment, & IDP: IELTS Australia. https://www.ielts.org/teaching-and-research/research-reports/online-series-2014-3

Wood, R. (1993). *Assessment and testing*. Cambridge University Press.

Wu, Y. (1998). What do tests of listening comprehension test? - A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*(1), 21-44. https://doi.org/10.1177/026553229801500102

Zarrabi, Z. (2020). Investigating the relationship between learning style and metacognitive listening awareness. *International Journal of Listening, 34*(1), 21-33. DOI: 10.1080/10904018.2016.1276458

**Supplementary Table 1**
**Identification of content words and AOI classification per test item**

Legend:
Yellow – Nouns
Orange – Verbs
Green – Adjectives
A = Answer
Q = Question

| Q1 | WC | A1 | WC | A2 | WC | A3 | WC | A4 | WC |
|---|---|---|---|---|---|---|---|---|---|
| Why | interrogative adverb | to | to-infinitive | to | to-infinitive | to | to-infinitive | to | to-infinitive |
| does | auxiliary verb | focus | verb | draw | verb | emphasize | verb | highlight | verb |
| the | determiner | on | preposition | a | determiner | the | determiner | the | determiner |
| instructor | noun | the | determiner | neutral | adjective | beauty | noun | limits | noun |
| compare | verb | value | noun | conclusion | noun | of | preposition | of | preposition |
| economic | adjective | of | preposition | about | preposition | economic | adjective | economic | adjective |
| models | noun | fashion | noun | fashion | noun | models | noun | models | noun |
| to | preposition | models | noun | models | noun | | | | |
| fashion | noun | | | | | | | | |
| models | noun | | | | | | | | |

| Q2 | WC | A1 | WC | A2 | WC | A3 | WC | A4 | WC |
|---|---|---|---|---|---|---|---|---|---|
| Why | interrogative adverb | They | pronoun | They | pronoun | They | pronoun | They | pronoun |
| are | auxiliary verb | are | auxiliary verb | supplement | verb | make | verb | showcase | verb |
| economic | adjective | replaceable | adjective | each | | the | determiner | important | adjective |
| models | noun | with | preposition | other | pronoun | discipline | noun | economic | adjective |
| regarded | verb | other | adjective | | | more | comparative adverb | activities | noun |
| as | preposition | models | noun | | | profitable | adjective | | |
| the | determiner | | | | | | | | |
| building | noun | | | | | | | | |
| blocks | noun | | | | | | | | |
| of | preposition | | | | | | | | |
| economics | noun | | | | | | | | |

| Q3 | WC | A1 | WC | A2 | WC | A3 | WC | A4 | WC |
|---|---|---|---|---|---|---|---|---|---|
| Which | interrogative adverb | investment | noun | decrease | noun | increased | adjective | impact | noun |
| of | preposition | return | noun | of | preposition | sale | noun | of | preposition |
| the | determiner | from | preposition | product | noun | of | preposition | earthquake | noun |
| could | | a | determiner | output | noun | new | adjective | on | preposition |
| be | auxiliary verb | foreign | adjective | after | preposition | model | noun | the | determiner |
| explained | verb | stock | noun | resource | noun | with | preposition | crop | noun |
| with | preposition | market | noun | re-allocation | noun | new | adjective | yields | noun |
| the | determiner | | | | | functions | noun | | |
| Production | noun | | | | | | | | |
| Possibility | noun | | | | | | | | |
| Curve | noun | | | | | | | | |

| Q4 | WC | A1 | WC | A2 | WC | A3 | WC | A4 | WC |
|---|---|---|---|---|---|---|---|---|---|
| What | interrogative adverb | visual | adjective | simple | adjective | Cobb-Douglas | noun | principle | adjective |
| is | auxiliary verb | and | conjunction | and | conjunction | and | conjunction | and | conjunction |
| one | determiner | mathematical | adjective | complex | adjective | Heckscher-Ohlin | noun | sophisticated | adjective |
| way | noun | | | | | | | | |
| of | preposition | | | | | | | | |
| grouping | noun | | | | | | | | |
| economic | adjective | | | | | | | | |
| models | noun | | | | | | | | |
| that | pronoun | | | | | | | | |
| the | determiner | | | | | | | | |
| professor | noun | | | | | | | | |
| mentions | verb | | | | | | | | |

| Q5 | WC | A1 | WC | A2 | WC | A3 | WC | A4 | WC |
|---|---|---|---|---|---|---|---|---|---|
| Fill | verb | mathematical | adjective | complex | adjective | graphic | adjective | simple | adjective |
| in | preposition | | | | | | | | |
| the | determiner | | | | | | | | |
| blank | noun | | | | | | | | |
| with | preposition | | | | | | | | |
| one | determiner | | | | | | | | |
| word | noun | | | | | | | | |
| from | preposition | | | | | | | | |
| the | determiner | | | | | | | | |
| lecture | noun | | | | | | | | |
| | | | | | | | | | |
| A | determiner | | | | | | | | |
| diagram | noun | | | | | | | | |
| is | auxiliary verb | | | | | | | | |
| a | determiner | | | | | | | | |
| type | noun | | | | | | | | |
| of | preposition | | | | | | | | |
| model | noun | | | | | | | | |

| Q6 | WC | A1 | WC | A2 | WC | A3 | WC | A4 | WC |
|---|---|---|---|---|---|---|---|---|---|
| The | determiner | a | determiner | a | determiner | a | determiner | an | determiner |
| instructor | noun | common | adjective | neglected | adjective | decisive | adjective | outcome | noun |
| mentions | verb | factor | noun | factor | noun | factor | noun | factor | noun |
| cultural | adjective | | | | | | | | |
| impact | noun | | | | | | | | |
| on | preposition | | | | | | | | |
| consumers' | noun | | | | | | | | |
| behaviour | noun | | | | | | | | |
| as | preposition | | | | | | | | |
| what | determiner | | | | | | | | |
| kind | noun | | | | | | | | |
| of | preposition | | | | | | | | |
| factor | noun | | | | | | | | |
| in | preposition | | | | | | | | |
| modelling | verb | | | | | | | | |
| economic | adjective | | | | | | | | |
| activities | noun | | | | | | | | |

| Q7 | WC | A1 | WC | A2 | WC | A3 | WC | A4 | WC |
|---|---|---|---|---|---|---|---|---|---|
| What | interrogative adverb | the | determiner | the | determiner | the | determiner | the | determiner |
| is | auxiliary verb | experience | noun | types | noun | history | noun | application | noun |
| the | determiner | of | preposition | of | preposition | of | preposition | of | preposition |
| one-size-fits-all | adjective | economists | noun | models | noun | economics | noun | models | noun |
| issue | noun | | | | | | | | |
| in | preposition | | | | | | | | |
| economic | adjective | | | | | | | | |
| modeling | noun | | | | | | | | |
| concerned | verb | | | | | | | | |
| with | preposition | | | | | | | | |

| Q8 | Wc | A1 | WC | A2 | WC | A3 | WC | A4 | WC |
|---|---|---|---|---|---|---|---|---|---|
| The | determiner | fixed | adjective | excluded | adjective | diminished | adjective | weighted | adjective |
| Latin | adjective | | | | | | | | |
| phrase | noun | | | | | | | | |
| ceteris paribus | noun | | | | | | | | |
| refers | verb | | | | | | | | |
| to | preposition | | | | | | | | |
| a | determiner | | | | | | | | |
| situation | noun | | | | | | | | |
| with | preposition | | | | | | | | |
| some | determiner | | | | | | | | |
| variables | noun | | | | | | | | |
| being | auxiliary verb | | | | | | | | |

| Q9 | WC | A1 | WC | A2 | WC | A3 | WC | A4 | WC |
|---|---|---|---|---|---|---|---|---|---|
| Which | interrogative adverb | Some | determiner | Economic | adjective | Economic | adjective | Some | determiner |
| of | preposition | models | noun | models | noun | models | noun | economic | adjective |
| the | determiner | are | auxiliary verb | always | adverb | should | | models | noun |
| following | adjective | superior | adjective | fall | | be | auxiliary verb | are | auxiliary verb |
| statements | noun | to | preposition | short | | as | | NOT | adverb |
| would | auxiliary verb | others | pronoun | in | verb | rigid | | solid | adjective |
| Dani Rodrik | noun | in | preposition | explaining | verb | as | adjective | in | preposition |
| the | determiner | predicting | verb | an | determiner | those | pronoun | theoretical | adjective |
| Harvard | noun | the | determiner | economy | noun | used | verb | foundation | noun |
| professor | noun | future | noun | | | in | preposition | | |
| probably | adverb | | | | | biology | noun | | |
| agree | | | | | | | | | |
| with | verb | | | | | | | | |

| Q10 | WC | A1 | WC | A2 | WC | A3 | WC | A4 | WC |
|---|---|---|---|---|---|---|---|---|---|
| Why | interrogative adverb | Real | adjective | Real | adjective | There | adverb | Modelling | verb |
| is | auxiliary verb | life | noun | life | noun | are | auxiliary verb | real | adjective |
| any | determiner | cannot | | would | auxiliary verb | too | adverb | life | noun |
| economic | adjective | be | auxiliary verb | require | verb | many | determiner | would | auxiliary verb |
| model | noun | adequately | adverb | sophisticated | adjective | variables | noun | take | verb |
| always | adverb | graphed | verb | mathematics | noun | in | preposition | too | adverb |
| simpler | adjective | | | | | real | adjective | much | determiner |
| than | preposition | | | | | life | noun | time | noun |
| real | adjective | | | | | | | | |
| life | noun | | | | | | | | |

| Q11 | WC | A1 | WC | A2 | WC | A3 | WC | A4 | WC | A5 | WC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Which | interrogative adverb | academic | adjective | the | determiner | the | determiner | current | adjective | an | determiner |
| of | preposition | achievement | noun | extent | noun | impact | noun | trends | noun | investigation | noun |
| the | determiner | across | preposition | of | preposition | of | preposition | in | preposition | into | preposition |
| following | pronoun | generations | noun | trade | noun | politics | noun | property | noun | public | adjective |
| could | | | | between | preposition | on | preposition | ownership | noun | transportation | noun |
| be | auxiliary verb | | | two | determiner | resource | noun | | | issues | noun |
| applied | verb | | | nations | noun | allocation | noun | | | | |
| to | preposition | | | | | | | | | | |
| an | determiner | | | | | | | | | | |
| economic | adjective | | | | | | | | | | |
| model | noun | | | | | | | | | | |
| Choose | verb | | | | | | | | | | |
| 2 | pronoun | | | | | | | | | | |

## Supplementary Table 2
## Results from Procedure 4 – Individual Content Word

| Test item | Classes of Content Words | Eye-tracking Variable | Content Word | Omnibus significance (*p* value) | Incorrect M (SD) | Partially Correct M (SD) | Correct M (SD) | Number of included cases *N* (%) |
|---|---|---|---|---|---|---|---|---|
| 1 | Nouns | Fixation Duration | A4_limits | .037 | 0.204 (0.563) | - | 0.270 (0.146) | 43 (65%) |
| | | Total Fixation Duration | Q_models | .029 | 0.603 (0.477) | - | 0.981 (0.715) | 50 (76%) |

| Test item | Classes of Content Words | Eye-tracking Variable | Content Word | Omnibus significance (*p* value) | Incorrect M (SD) | Partially Correct M (SD) | Correct M (SD) | Number of included cases *N* (%) |
|---|---|---|---|---|---|---|---|---|
| | | Fixation Count | Q_models | .021 | 2.52 (1.78) | - | 4.10 (2.85) | 50 (76%) |
| | | Total Visit Duration | Q_models | .032 | 0.608 (0.482) | - | 0.982 (0.718) | 50 (76%) |
| | | Visit Count | Q_models | .007 | 2.29 (1.55) | - | 4.03 (2.72) | 50 (76%) |
| | Verbs | Fixation Duration | A4_highlight | .004 | 0.224 (0.0528) | - | 0.288 (0.100) | 47 (71%) |
| 2 | Nouns | Fixation Duration | Q_building | .013 | 0.271 (0.0571) | - | 0.223 (0.0518) | 49 (74%) |
| | | Visit Duration | Q_building | .002 | 0.322 (0.0851) | - | 0. 237 (0.0656) | 49 (74%) |
| 3 | Nouns | Fixation Duration | A2_resource | .044 | 0.223 (0.0703) | - | 0.272 (0.0893) | 54 (82%) |
| | | Total Fixation Duration | A2_resource | .036 | 0.528 (0.371) | - | 0.826 (0.511) | 54 (82%) |
| | | | A3_functions | .018 | 0.317 (0.162) | - | 0.596 (0.499) | 37 (56%) |
| | | | A4_impact | .016 | 1.19 (1.08) | - | 0.628 (0.323) | 42 (64%) |
| | | Fixation Count | A4_impact | .042 | 4.23 (3.06) | - | 2.72 (1.51) | 42 (64%) |
| | | Visit Duration | A2_resource | .007 | 0.223 (0.0703) | - | 0.302 (0.133) | 54 (82%) |
| | | Total Visit Duration | A2_resource | .033 | 0.528 (0.371) | - | 0.847 (0.557) | 54 (82%) |
| | | | A3_functions | .018 | 0.317 (0.162) | - | 0.630 (0.619) | 37 (56%) |
| | | | A4_impact | .028 | 1.20 (1.09) | - | 0.667 (0.380) | 42 (64%) |
| | | Visit Count | A4_impact | .033 | 4.00 (2.77) | - | 2.59 (1.32) | 42 (64%) |
| | Adjectives | Total Fixation Duration | A3_new2 | .026 | 0.323 (0.117) | - | 0.514 (0.324) | 44 (67%) |
| | | Total Visit Duration | A3_new2 | .026 | 0.323 (0.117) | - | 0.514 (0.324) | 44 (67%) |
| 4 | Nouns | Total Fixation | Q_way | .013 | 0.282 (0.0975) | - | 0.612 (0.483) | 52 (79%) |

| Test item | Classes of Content Words | Eye-tracking Variable | Content Word | Omnibus significance (*p* value) | Incorrect M (SD) | Partially Correct M (SD) | Correct M (SD) | Number of included cases *N* (%) |
|---|---|---|---|---|---|---|---|---|
| | | Duration | Q_models | .043 | 0.354 (0.204) | - | 0.843 (0.721) | 38 (58%) |
| | | Fixation Count | Q_way | .010 | 1.33 (0.516) | - | 2.74 (2.14) | 52 (79%) |
| | | | Q_models | .032 | 1.60 (0.894) | - | 3.76 (3.08) | 38 (58%) |
| | | Total Visit Duration | Q_way | .015 | 0.285 (0.101) | - | 0.619 (0.497) | 52 (79%) |
| | | Visit Count | Q_way | .002 | 1.17 (0.408) | - | 2.72 (2.14) | 52 (79%) |
| | | | Q_models | .016 | 1.40 (0.894) | - | 3.64 (2.97) | 38 (58%) |
| | Adjectives | Fixation Duration | A4_sophisticated | .007 | 0.372 (0.131) | - | 0.234 (0.0931) | 32 (49%) |
| | | Total Fixation Duration | Q_economic | .036 | 0.596 (0.531) | - | 1.30 (1.08) | 47 (71%) |
| | | Fixation Count | Q_economic | .040 | 2.63 (2.45) | - | 5.51 (4.54) | 47 (71%) |
| | | Visit Duration | A4_sophisticated | .002 | 0.413 (0.132) | - | 0.252 (0.0926) | 32 (49%) |
| | | Total Visit Duration | Q_economic | .036 | 0.598 (0.532) | - | 1.30 (1.08) | 47 (71%) |
| | | Visit Count | Q_economic | .036 | 2.50 (2.39) | - | 5.28 (4.32) | 47 (71%) |
| | | | A1_visual | .042 | 2.38 (1.77) | - | 4.41 (3.99) | 54 (82%) |
| 5 | Adjectives | Visit Duration | A4_simple (OP) | .019 | 0.508 (0.198) | - | 0.332 (0.160) | 26 (39%) |
| 6 | Verbs | Fixation Duration | Q_modeling (CL Ans graphic) | .001 | 0.183 (0.0497) | - | 0.287 (0.0848) | 24 (36%) |
| | | Visit Duration | Q_modeling (CL Ans graphic) | .001 | 0.187 (0.0524) | - | 0.305 (0.0988) | 24 (36%) |
| | Adjectives | Fixation Duration | Q_cultural (CL Ans graphic) | .004 | 0.160 (0.0592) | - | 0.245 (0.0633) | 21 (32%) |
| | | Total Fixation Duration | Q_economic (CL) | .030 | 0.04 (1.55) | - | 0.818 (0.837) | 30 (46%) |
| | | Total Visit Duration | Q_economic (CL) | .031 | 2.06 (1.56) | - | 0.826 (0.856) | 30 (46%) |

| Test item | Classes of Content Words | Eye-tracking Variable | Content Word | Omnibus significance (*p* value) | Incorrect M (SD) | Partially Correct M (SD) | Correct M (SD) | Number of included cases *N* (%) |
|---|---|---|---|---|---|---|---|---|
| 7 | Nouns | Visit Duration | A1_experience (CL) | .011 | 1.20 (1.38) | - | 0.296 (0.122) | 33 (50%) |
| | | Visit Count | A1_experience (CL) | .040 | 1.67 (1.16) | - | 4.40 (2.79) | 33 (50%) |
| 8 | Adjectives | Fixation Duration | A3_diminished (OP) | .042 | 0.558 (0.399) | - | 0.336 (0.165) | 25 (38%) |
| | | Total Fixation Duration | A1_fixed (OP) | .049 | 0.938 (0.604) | - | 0.550 (0.322) | 23 (35%) |
| | | Visit Duration | A3_diminished (OP) | .016 | 0.608 (0.388) | - | 0.346 (0.175) | 25 (38%) |
| | | Total Visit Duration | A1_fixed (OP) | .050 | 0.941 (0.604) | - | 0.552 (0.328) | 23 (35%) |
| | | | A3_diminished (OP) | .045 | 1.27 (0.729) | - | 0.747 (0.565) | 25 (38%) |
| | | Visit Count | A1_fixed (OP) | .045 | 2.80 (1.87) | - | 1.62 (0.961) | 23 (35%) |
| 11 | Nouns | Visit Duration | A3_impact | .037 | 0.460 (0.225) | 0.264 (0.0792) | 0.244 (0.0783) | 20 (30%) |

Note. *Refer to Table 3 in the Supplementary File for significant p values with fewer than 20 (30%) included cases. A = Answer; Q = Question.*

## Supplementary Table 3
## Results from Procedure 4 for significant *p* values with fewer than 20 (30.3%) included cases

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| 1 | Nouns | Fixation Duration | Q_models2 | 0.054 | 0.318 (0.151) | 0.210 (0.0616) | 15 (22.7%) |
| | | | A2_fashion | 0.066 | 0.203 (0.635) | 0.247 (0.0671) | 30 (45.5%) |
| | | Visit Duration | Q_models2 | 0.054 | 0.318 (0.151) | 0.210 (0.0616) | 15 (22.7%) |
| | | | A4_limits | 0.074 | 0.218 (0.0672) | 0.279 (0.148) | 43 (65.2%) |
| | Verbs | Visit Duration | A4_highlight | 0.089 | 0.257 (0.0770) | 0.302 (0.105) | 47 (71.2%) |
| | Adjectives | Fixation Count | A2_neutral | 0.078 | 3.88 (2.86) | 2.78 (1.76) | 56 (84.8%) |
| | | Visit Count | A2_neutral | 0.068 | 3.75 (2.86) | 2.66 (1.56) | 56 (84.8%) |
| 2 | Nouns | Fixation Duration | A4_activities | 0.08 | 0.168 (0.0492) | 0.222 (0.0589) | 17 (25.8%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| | | Total Fixation Duration | A4_activities | 0.085 | 0.228 (0.164) | 0.477 (0.332) | 17 (25.8%) |
| | | Visit Duration | A4_activities | 0.032 | 0.168 (0.0492) | 0.235 (0.0549) | 17 (25.8%) |
| | | Total Visit Duration | A4_activities | 0.083 | 0.228 (0.164) | 0.480 (0.333) | 17 (25.8%) |
| | Adjectives | Visit Duration | Q_economic | 0.072 | 0.304 (0.170) | 0.246 (0.0589) | 60 (90.9%) |
| 3 | Nouns | Total Fixation Duration | A3_model | 0.087 | 0.328 (0.178) | 0.514 (0.401) | 44 (66.7%) |
| | | Fixation Count | A3_functions | 0.052 | 1.42 (0.669) | 2.16 (1.52) | 37 (56.1%) |
| | | Visit Duration | A3_functions | 0.086 | 0.227 (0.0593) | 0.304 (0.190) | 37 (56.1%) |
| | | Total Visit Duration | A3_model | 0.087 | 0.328 (0.178) | 0.514 (0.401) | 44 (66.7%) |
| | | Visit Count | A3_functions | 0.069 | 1.42 (0.669) | 1.92 (0.862) | 37 (56.1%) |
| | Adjectives | Fixation Count | A3_new2 | 0.072 | 1.42 (0.515) | 1.91 (0.963) | 44 (66.7%) |
| | | Visit Count | A3_new2 | 0.072 | 1.42 (0.515) | 1.91 (0.963) | 44 (66.7%) |
| 4 | Nouns | Visit Duration | Q_grouping | 0.075 | 0.208 (0.0514) | 0.249 (0.0693) | 53 (80.3%) |
| | | Total Visit Duration | Q_models | 0.067 | 0.388 (0.217) | 0.850 (0.725) | 38 (57.6%) |
| | Adjectives | Fixation Duration | A2_complex | 0.077 | 0.278 (0.0606) | 0.222 (0.0660) | 48 (72.7%) |
| | | Fixation Count | A1_visual | 0.059 | 2.50 (1.85) | 4.57 (4.33) | 54 (81.8%) |
| 5 | Nouns | Fixation Duration | Q_word (CL Ans mathematical) | 0.02 | 0.264 (0.0793) | 0.0400 (-) | 6 (9.1%) |
| | | Total Fixation Duration | Q_blank (CL Ans graphic) | 0.014 | 1.00 (-) | 0.387 (0.287) | 8 (12.1%) |
| | | | Q_word (CL Ans graphic) | 0.007 | 3.47 (-) | 0.406 (0.226) | 14 (21.2%) |
| | | | Q_type (CL Ans mathematical) | 0.051 | 0.285 (0.0778) | 0.210 (-) | 3 (4.5%) |
| | | | Q_word (CL Ans mathematical) | 0.02 | 0.426 (0.218) | 0.0400 (-) | 6 (9.1%) |
| | | Fixation Count | Q_blank (CL Ans graphic) | 0.014 | 6.00 (-) | 2.14 (1.22) | 8 (12.1%) |
| | | | Q_word (CL Ans graphic) | 0.007 | 11.00 (-) | 1.77 (0.725) | 14 (21.2%) |
| | | Visit Duration | Q_word (CL Ans graphic) | 0.063 | 0.380 (-) | 0.223 (0.0745) | 14 (21.2%) |
| | | | Q_word | 0.02 | 0.264 (0.0793) | 0.0400 (-) | 6 (9.1%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| | | | (CL Ans mathematical) | | | | |
| | | Total Visit Duration | Q_blank (CL Ans graphic) | 0.014 | 1.00 (-) | 0.387 (0.287) | 8 (12.1%) |
| | | | Q_word (CL Ans graphic) | 0.007 | 3.84 (-) | 0.406 (0.226) | 14 (21.2%) |
| | | | Q_type (CL Ans mathematical) | 0.051 | 0.285 (0.0778) | 0.210 (-) | 3 (4.5%) |
| | | | Q_word (CL Ans mathematical) | 0.02 | 0.426 (0.218) | 0.0400 (-) | 6 (9.1%) |
| | | Visit Count | Q_blank (CL Ans graphic) | 0.014 | 6.00 (-) | 2.14 (1.22) | 8 (12.1%) |
| | | | Q_word (CL Ans graphic) | 0.007 | 10.00 (-) | 1.77 (0.725) | 14 (21.2%) |
| | Verbs | Fixation Duration | Q_fill (CL Ans graphic) | 0.051 | 0.170 (-) | 0.485 (0.0212) | 3 (4.5%) |
| | | | Q_fill (OP) | 0.077 | 0.225 (0.0919) | 0.363 (0.0981) | 6 (9.1%) |
| | | Total Fixation Duration | Q_fill (CL Ans graphic) | 0.051 | 0.170 (-) | 0.485 (0.0212) | 3 (4.5%) |
| | | | Q_fill (OP) | 0.077 | 0.225 (0.0919) | 0.363 (0.0981) | 6 (9.1%) |
| | | Visit Duration | Q_fill (CL Ans graphic) | 0.051 | 0.170 (-) | 0.485 (0.0212) | 3 (4.5%) |
| | | | Q_fill (OP) | 0.077 | 0.225 (0.0919) | 0.363 (0.0981) | 6 (9.1%) |
| | | Total Visit Duration | Q_fill (CL Ans graphic) | 0.051 | 0.170 (-) | 0.485 (0.0212) | 3 (4.5%) |
| | | | Q_fill (OP) | 0.077 | 0.225 (0.0919) | 0.363 (0.0981) | 6 (9.1%) |
| | Adjectives | Fixation Duration | A1_mathematical (OP Ans graphic) | 0.02 | 1.02 (-) | 0.230 (0.0822) | 6 (9.1%) |
| | | Visit Duration | A1_mathematical (OP Ans graphic) | 0.02 | 1.02 (-) | 0.246 (0.0918) | 6 (9.1%) |
| 6 | Nouns | Fixation Duration | Q_instructor (CL Ans mathematical) | 0.018 | 0.260 (0.0141) | 0.194 (0.0427) | 10 (15.2%) |
| | | | Q_impact (CL Ans mathematical) | 0.012 | 0.400 (-) | 0.264 (0.0661) | 9 (13.6%) |
| | | | Q_activities (CL | 0.053 | 0.250 (-) | 0.210 (0.0260) | 10 (15.2%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| | | | Ans mathematical) | | | | |
| | | | A1_factor (CL Ans graphic) | 0.026 | 0.163 (0.0206) | 0.206 (0.0436) | 18 (27.3%) |
| | | | A4_outcome (CL Ans mathematical) | 0.061 | 0.120 (-) | 0.335 (0.153) | 9 (13.6%) |
| | | Total Fixation Duration | Q_consumers' (CL Ans graphic) | 0.035 | 0.452 (0.559) | 1.11 (0.675) | 18 (27.3%) |
| | | | Q_behaviour (CL Ans graphic) | 0.031 | 0.340 (0.212) | 0.986 (0.733) | 16 (24.2%) |
| | | | Q_kind (CL Ans mathematical) | 0.067 | 0.920 (0.523) | 0.391 (0.292) | 10 (15.2%) |
| | | | A3_factor (CL) | 0.016 | 0.518 (0.545) | 0.199 (0.107) | 17 (25.8%) |
| | | | A4_outcome (CL Ans graphic) | 0.088 | 1.04 (0.770) | 0.612 (0.424) | 25 (37.9%) |
| | | | A4_outcome (CL Ans mathematical) | 0.061 | 0.120 (-) | 0.711 (0.392) | 9 (13.6%) |
| | | Fixation Count | Q_consumers' (CL Ans graphic) | 0.064 | 2.17 (2.86) | 4.92 (3.18) | 18 (27.3%) |
| | | | Q_behaviour (CL Ans graphic) | 0.039 | 1.80 (1.30) | 4.91 (3.59) | 16 (24.2%) |
| | | | Q_impact (CL Ans mathematical) | 0.061 | 1.00 (-) | 4.00 (3.21) | 9 (13.6%) |
| | | | Q_kind (CL Ans mathematical) | 0.019 | 3.50 (2.12) | 1.38 (0.518) | 10 (15.2%) |
| | | | Q_factor (CL Ans mathematical) | 0.067 | 1.50 (0.707) | 3.67 (1.75) | 8 (12.1%) |
| | | | A3_factor (CL) | 0.027 | 2.00 (1.10) | 1.18 (0.405) | 17 (25.8%) |
| | | Visit Duration | Q_impact (CL Ans mathematical) | 0.012 | 0.400 (-) | 0.264 (0.0661) | 9 (13.6%) |
| | | | Q_activities (CL Ans mathematical) | 0.053 | 0.250 (-) | 0.210 (0.0260) | 10 (15.2%) |
| | | | A1_factor (CL Ans graphic) | 0.026 | 0.163 (0.0206) | 0.206 (0.0436) | 18 (27.3%) |
| | | | A4_outcome (CL | 0.061 | 0.120 (-) | 0.335 (0.153) | 9 (13.6%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| | | | Ans mathematical) | | | | |
| | | Total Visit Duration | Q_consumers' (CL Ans graphic) | 0.031 | 0.453 (0.563) | 1.17 (0.733) | 18 (27.3%) |
| | | | Q_behaviour (CL Ans graphic) | 0.031 | 0.340 (0.212) | 1.01 (0.769) | 16 (24.2%) |
| | | | Q_kind (CL Ans mathematical) | 0.067 | 0.920 (0.523) | 0.391 (0.292) | 10 (15.2%) |
| | | | A3_factor (CL) | 0.016 | 0.518 (0.545) | 0.199 (0.107) | 17 (25.8%) |
| | | | A4_outcome (CL Ans mathematical) | 0.061 | 0.120 (-) | 0.711 (0.392) | 9 (13.6%) |
| | | Visit Count | Q_consumers' (CL Ans graphic) | 0.064 | 2.00 (2.45) | 4.58 (3.18) | 18 (27.3%) |
| | | | Q_behaviour (CL Ans graphic) | 0.045 | 1.80 (1.30) | 4.64 (3.36) | 16 (24.2%) |
| | | | Q_impact (CL Ans mathematical) | 0.061 | 1.00 (-) | 4.00 (3.21) | 9 (13.6%) |
| | | | Q_kind (CL Ans mathematical) | 0.019 | 3.50 (2.12) | 1.38 (0.518) | 10 (15.2%) |
| | | | A3_factor (CL) | 0.027 | 2.00 (1.10) | 1.18 (0.405) | 17 (25.8%) |
| | | | A1_factor (CL Ans mathematical) | 0.069 | - (-) | 1.40 (0.548) | 8 (12.1%) |
| | Verbs | Visit Count | Q_mentions (CL Ans mathematical) | 0.085 | 1.33 (0.577) | 3.75 (3.33) | 11 (16.7%) |
| | Adjectives | Total Fixation Duration | Q_cultural (CL Ans mathematical) | 0.012 | 0.330 (-) | 2.04 (1.77) | 9 (13.6%) |
| | | Fixation Count | Q_economic (CL) | 0.077 | 7.20 (4.21) | 3.60 (3.69) | 30 (45.5%) |
| | | | Q_cultural (CL Ans mathematical) | 0.061 | 1.00 (-) | 6.63 (5.34) | 9 (13.6%) |
| | | | A2_neglected (CL Ans graphic) | 0.084 | 9.14 (5.34) | 6.10 (3.24) | 27 (40.9%) |
| | | Visit Duration | Q_cultural (CL Ans graphic) | 0.054 | 0.184 (0.0820) | 0.258 (0.0811) | 21 (31.8%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| | | | A1_common (CL Ans graphic) | 0.051 | 0.239 (0.0398) | 0.330 (0.148) | 26 (39.4%) |
| | | Total Visit Duration | Q_cultural (CL Ans mathematical) | 0.012 | 0.330 (-) | 2.08 (1.78) | 9 (13.6%) |
| | | Visit Count | Q_economic (CL) | 0.078 | 6.00 (3.81) | 3.16 (2.82) | 30 (45.5%) |
| | | | Q_cultural (CL Ans mathematical) | 0.061 | 1.00 (-) | 5.75 (4.06) | 9 (13.6%) |
| 7 | Nouns | Fixation Duration | Q_modeling (CL Ans mathematical) | 0.051 | 0.230 (-) | 0.175 (0.0636) | 3 (4.5%) |
| | | | A2_models (CL Ans graphic) | 0.007 | 0.120 (-) | 0.206 (0.0453) | 15 (22.7%) |
| | | | A3_history (CL Ans graphic) | 0.006 | 0.170 (-) | 0.269 (0.0847) | 17 (25.8%) |
| | | Total Fixation Duration | Q_issue (CL Ans mathematical) | 0.02 | 0.770 (-) | 0.326 (0.148) | 6 (9.1%) |
| | | | Q_modeling (CL Ans graphic) | 0.033 | 0.185 (0.0212) | 0.657 (0.488) | 18 (27.3%) |
| | | | Q_modeling (CL Ans mathematical) | 0.051 | 0.700 (-) | 0.175 (0.0636) | 3 (4.5%) |
| | | | A1_experience (CL) | 0.093 | 0.560 (0.235) | 1.26 (0.970) | 33 (50.0%) |
| | | | A1_experience (CL Ans graphic) | 0.094 | 0.810 (0.0141) | 1.43 (0.602) | 24 (36.4%) |
| | | | A1_economists (CL Ans graphic) | 0.088 | 0.280 (0.171) | 0.660 (0.433) | 13 (19.7%) |
| | | | A2_models (CL Ans graphic) | 0.007 | 0.120 (-) | 0.559 (0.419) | 15 (22.7%) |
| | | | A3_economics (CL Ans graphic) | 0.077 | 0.170 (-) | 0.425 (0.216) | 18 (27.3%) |
| | | Fixation Count | Q_issue (CL Ans mathematical) | 0.02 | 3.00 (-) | 1.40 (0.548) | 6 (9.1%) |
| | | | Q_modeling (CL Ans graphic) | 0.041 | 1.00 (-) | 3.38 (2.47) | 18 (27.3%) |
| | | | Q_modeling (CL Ans mathematical) | 0.051 | 3.00 (-) | 1.00 (0.00) | 3 (4.5%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| | | | A1_experience (CL) | 0.051 | 2.00 (1.00) | 5.00 (3.44) | 33 (50.0%) |
| | | | A1_experience (CL Ans graphic) | 0.094 | 3.50 (0.707) | 6.05 (2.40) | 24 (36.4%) |
| | | | A1_economists (CL Ans graphic) | 0.034 | 1.00 (0.00) | 2.90 (1.91) | 13 (19.7%) |
| | | Visit Duration | Q_modeling (CL Ans mathematical) | 0.051 | 0.230 (-) | 0.175 (0.0636) | 3 (4.5%) |
| | | | A2_models (CL Ans graphic) | 0.007 | 0.120 (-) | 0.210 (0.0451) | 15 (22.7%) |
| | | | A3_history (CL Ans graphic) | 0.006 | 0.170 (-) | 0.271 (0.0841) | 17 (25.8%) |
| | | Total Visit Duration | Q_issue (CL Ans mathematical) | 0.02 | 0.770 (-) | 0.326 (0.148) | 6 (9.1%) |
| | | | Q_modeling (CL Ans graphic) | 0.033 | 0.185 (0.0212) | 0.659 (0.491) | 18 (27.3%) |
| | | | Q_modeling (CL Ans mathematical) | 0.051 | 0.700 (-) | 0.175 (0.0636) | 3 (4.5%) |
| | | | A1_economists (CL Ans graphic) | 0.087 | 0.280 (0.171) | 0.663 (0.437) | 13 (19.7%) |
| | | | A2_models (CL Ans graphic) | 0.007 | 0.120 (-) | 0.571 (0.441) | 15 (22.7%) |
| | | | A3_economics (CL Ans graphic) | 0.077 | 0.170 (-) | 0.425 (0.216) | 18 (27.3%) |
| | | Visit Count | Q_issue (CL Ans mathematical) | 0.02 | 3.00 (-) | 1.40 (0.548) | 6 (9.1%) |
| | | | Q_modeling (CL Ans mathematical) | 0.051 | 3.00 (-) | 1.00 (0.00) | 3 (4.5%) |
| | | | Q_modeling (CL Ans graphic) | 0.041 | 1.00 (0.00) | 3.25 (2.27) | 18 (27.3%) |
| | | | A1_experience (CL Ans graphic) | 0.076 | 3.00 (0.00) | 5.41 (2.06) | 24 (36.4%) |
| | | | A1_economists (CL Ans graphic) | 0.034 | 1.00 (0.00) | 2.70 (1.70) | 13 (19.7%) |
| | Adjectives | Fixation Duration | Q_one-size-fits-all (CL Ans mathematical) | 0.012 | 0.360 (-) | 0.230 (0.0556) | 9 (13.6%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| | | Total Fixation Duration | Q_economic (CL Ans mathematical) | 0.02 | 2.54 (-) | 0.428 (0.435) | 6 (9.1%) |
| | | Fixation Count | Q_economic (CL Ans mathematical) | 0.02 | 11.0 (-) | 2.00 (1.73) | 6 (9.1%) |
| | | Visit Duration | Q_one-size-fits-all (CL) | 0.092 | 0.193 (0.0116) | 0.279 (0.118) | 39 (59.1%) |
| | | | Q_economic (CL Ans mathematical) | 0.02 | 0.320 (-) | 0.192 (0.0654) | 6 (9.1%) |
| | | Total Visit Duration | Q_economic (CL Ans mathematical) | 0.02 | 2.60 (-) | 0.428 (0.435) | 6 (9.1%) |
| | | Visit Count | Q_economic (CL Ans mathematical) | 0.02 | 8.00 (-) | 2.00 (1.73) | 6 (9.1%) |
| 8 | Nouns | Fixation Duration | Q_ceteris paribus (CL) | 0.087 | 0.244 (0.101) | 0.195 (0.0519) | 30 (45.5%) |
| | | | Q_variables (OP) | 0.051 | 0.230 (0.0424) | 0.160 (-) | 3 (4.5%) |
| | | Total Fixation Duration | Q_phrase (OP) | 0.03 | 0.508 (0.319) | 0.214 (0.0901) | 9 (13.6%) |
| | | | Q_variables (OP) | 0.051 | 0.360 (0.226) | 0.160 (-) | 3 (4.5%) |
| | | Fixation Count | Q_phrase (OP) | 0.063 | 2.00 (0.816) | 1.20 (0.447) | 9 (13.6%) |
| | | Visit Duration | Q_ceteris paribus (CL) | 0.099 | 0.299 (0.174) | 0.221 (0.0809) | 30 (45.5%) |
| | | | Q_variables (OP) | 0.051 | 0.230 (0.0424) | 0.160 (-) | 3 (4.5%) |
| | | Total Visit Duration | Q_phrase (OP) | 0.03 | 0.543 (0.385) | 0.214 (0.0902) | 9 (13.6%) |
| | | | Q_variables (OP) | 0.051 | 0.360 (0.226) | 0.160 (-) | 3 (4.5%) |
| | Adjectives | Total Fixation Duration | A3_diminished (OP) | 0.06 | 1.21 (0.674) | 0.746 (0.562) | 25 (37.9%) |
| | | Fixation Count | A1_fixed (OP) | 0.076 | 3.00 (1.89) | 1.77 (1.42) | 23 (34.8%) |
| 9 | Nouns | Total Fixation Duration | A3_models (CL Ans fixed) | 0.018 | 0.715 (0.177) | 0.333 (0.177) | 17 (25.8%) |
| | | | A4_models (CL Ans fixed) | 0.012 | 0.180 (-) | 0.423 (0.215) | 9 (13.6%) |
| | | Fixation Count | A2_models (CL) | 0.073 | 1.33 (0.516) | 2.30 (1.49) | 16 (24.2%) |
| | | | A3_models (CL) | 0.095 | 1.00 (0.00) | 1.63 (1.19) | 15 (22.7%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| | | | A3_models (CL Ans fixed) | 0.048 | 2.50 (0.707) | 1.47 (0.640) | 17 (25.8%) |
| | | Visit Duration | A1_models (CL) | 0.098 | 0.296 (0.162) | 0.223 (0.0736) | 27 (40.9%) |
| | | | A4_models (CL Ans fixed) | 0.012 | 0.180 (-) | 0.285 (0.0678) | 9 (13.6%) |
| | | Total Visit Duration | A3_models (CL Ans fixed) | 0.02 | 0.715 (0.177) | 0.343 (0.181) | 17 (25.8%) |
| | | | A4_models (CL Ans fixed) | 0.012 | 0.180 (-) | 0.425 (0.214) | 9 (13.6%) |
| | | Visit Count | A2_models (CL) | 0.073 | 1.33 (0.516) | 2.30 (1.49) | 16 (24.2%) |
| | | | A3_models (CL) | 0.095 | 1.00 (0.00) | 1.50 (0.926) | 15 (22.7%) |
| | | | A3_models (CL Ans fixed) | 0.041 | 2.50 (0.707) | 1.40 (0.632) | 17 (25.8%) |
| | Verbs | Fixation Duration | Q_agree (CL) | 0.085 | 0.267 (0.0984) | 0.204 (0.0648) | 21 (31.8%) |
| | | | A3_used (CL Ans fixed) | 0.014 | 0.130 (-) | 0.264 (0.0922) | 8 (12.1%) |
| | | Visit Duration | Q_agree (CL) | 0.085 | 0.267 (0.0984) | 0.204 (0.0648) | 21 (31.8%) |
| | | | A3_used (CL Ans fixed) | 0.071 | 0.430 (-) | 0.264 (0.0922) | 8 (12.1%) |
| | Adjectives | Fixation Duration | Q_following (CL Ans diminished) | 0.096 | 0.970 (-) | 0.140 (-) | 2 (3.0%) |
| | | Total Fixation Duration | A3_Economic (CL) | 0.096 | 0.235 (0.0495) | 0.518 (0.392) | 10 (15.2%) |
| | | | Q_following (CL Ans diminished) | 0.096 | 0.970 (-) | 0.140 (-) | 2 (3.0%) |
| | | Fixation Count | A3_Economic (CL) | 0.035 | 1.00 (0.00) | 2.50 (1.41) | 10 (15.2%) |
| | | | A4_solid (CL) | 0.013 | 1.67 (0.577) | 1.00 (0.00) | 10 (15.2%) |
| | | | A1_superior (CL Ans fixed) | 0.095 | 1.00 (0.00) | 1.67 (0.724) | 17 (25.8%) |
| | | Visit Duration | Q_following (CL Ans diminished) | 0.096 | 0.970 (-) | 0.140 (-) | 2 (3.0%) |
| | | Total Visit Duration | A3_Economic (CL) | 0.096 | 0.235 (0.0495) | 0.528 (0.415) | 10 (15.2%) |
| | | | Q_following (CL Ans diminished) | 0.096 | 0.970 (-) | 0.140 (-) | 2 (3.0%) |
| | | Visit Count | A3_Economic | 0.035 | 1.00 (0.00) | 2.25 (1.04) | 10 (15.2%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| | | | (CL) | | | | |
| | | | A4_solid (CL) | 0.013 | 1.00 (0.00) | 1.67 (0.724) | 10 (15.2%) |
| | | | A1_superior (CL Ans fixed) | 0.095 | 1.67 (0.577) | 1.00 (0.00) | 17 (25.8%) |
| 10 | Nouns | Fixation Duration | Q_life (CL) | 0.034 | 0.220 (-) | 0.153 (0.0252) | 4 (6.1%) |
| | | Total Fixation Duration | A3_life (CL) | 0.009 | 0.310 (0.0283) | 0.187 (0.0702) | 5 (7.6%) |
| | | | A3_variables (CL) | 0.042 | 0.180 (0.0529) | 0.619 (0.592) | 19 (28.2%) |
| | | | A1_life (CL Ans fixed) | 0.034 | 0.460 (-) | 0.263 (0.0833) | 4 (6.1%) |
| | | Fixation Count | A3_variables (CL) | 0.024 | 1.00 (0.00) | 2.81 (2.61) | 19 (28.8.%) |
| | | | A1_life (CL Ans fixed) | 0.034 | 2.00 (-) | 1.00 (0.00) | 4 (6.1%) |
| | | Visit Duration | Q_life (CL) | 0.034 | 0.220 (-) | 0.153 (0.0252) | 4 (6.1%) |
| | | Total Visit Duration | A3_life (CL) | 0.009 | 0.310 (0.0283) | 0.187 (0.0702) | 5 (7.6%) |
| | | | A3_variables (CL) | 0.04 | 0.180 (0.0529) | 0.634 (0.621) | 19 (28.8%) |
| | | | A1_life (CL Ans fixed) | 0.034 | 0.460 (-) | 0.263 (0.0833) | 4 (6.1%) |
| | | Visit Count | A3_variables (CL) | 0.037 | 1.00 (0.00) | 2.56 (2.19) | 19 (28.8%) |
| | Verbs | Fixation Duration | A4_take (CL) | 0.017 | 0.110 (-) | 0.250 (0.123) | 7 (10.6%) |
| | | | A4_Modeling (CL Ans fixed) | 0.051 | 0.190 (-) | 0.210 (0.0141) | 3 (4.5%) |
| | | Total Fixation Duration | A4_take (CL) | 0.017 | 0.110 (-) | 0.348 (0.291) | 7 (10.6%) |
| | | | A4_Modeling (CL Ans fixed) | 0.051 | 0.190 (-) | 0.210 (0.0141) | 3 (4.5%) |
| | | Visit Duration | A4_take (CL) | 0.017 | 0.110 (-) | 0.250 (0.123) | 7 (10.6%) |
| | | | A4_Modeling (CL Ans fixed) | 0.051 | 0.190 (-) | 0.210 (0.0141) | 3 (4.5%) |
| | | Total Visit Duration | A4_take (CL) | 0.017 | 0.110 (-) | 0.348 (0.291) | 7 (10.6%) |
| | | | A4_Modeling (CL Ans fixed) | 0.051 | 0.190 (-) | 0.210 (0.0141) | 3 (4.5%) |
| | Adjectives | Total Fixation Duration | Q_simpler (CL) | 0.032 | 0.249 (0.130) | 0.521 (0.315) | 19 (28.8%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|
| | | Fixation Count | Q_simpler (CL) | 0.038 | 1.29 (0.488) | 2.42 (1.38) | 19 (28.8%) |
| | | Total Visit Duration | Q_simpler (CL) | 0.034 | 0.251 (0.132) | 0.523 (0.317) | 19 (28.8%) |
| | | Visit Count | Q_simpler (CL) | 0.021 | 1.14 (0.378) | 2.33 (1.30) | 19 (28.8%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | μ 1 (SD) | μ 2 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|---|
| 11 | Nouns | Fixation Duration | A1_generations | 0.081 | 2.33 (1.16) | 1.83 (1.27) | 4.17 (3.31) | 21 (31.8%) |
| | | | A2_trade | 0.084 | 0.225 (0.00707) | 0.186 (0.0505) | 0.247 (0.0628) | 17 (25.8%) |
| | | | A3_allocation | 0.051 | - (-) | 0.220 (-) | 0.295 (0.0778) | 3 (4.55%) |
| | | | A4_property | 0.033 | 0.153 (0.0306) | 0.225 (0.0707) | 0.163 (0.0250) | 15 (22.7%) |
| | | | A5_investigation | 0.04 | 0.330 (0.0849) | 0.220 (0.0738) | 0.168 (0.0929) | 15 (22.7%) |
| | | Total Fixation Duration | A2_nations | 0.049 | 0.510 (-) | 0.208 (0.0699) | 0.185 (0.00707) | 7 (10.6%) |
| | | | A3_allocation | 0.051 | - (-) | 0.220 (-) | 0.295 (0.0778) | 3 (4.55%) |
| | | | A5_transportation | 0.084 | 0.258 (0.0922) | 0.411 (0.304) | 0.637 (0.197) | 14 (21.2%) |
| | | Fixation Count | A2_nations | 0.057 | 2.00 (-) | 1.00 (0.00) | 1.00 (0.00) | 7 (10.6%) |
| | | | A5_transportation | 0.025 | 1.00 (0.00) | 2.00 (1.41) | 2.33 (0.577) | 14 (21.2%) |
| | | Visit Duration | A2_trade | 0.084 | 0.225 (0.00707) | 0.186 (0.0505) | 0.247 (0.0628) | 17 (25.8%) |
| | | | A3_allocation | 0.051 | - (-) | 0.220 (-) | 0.295 (0.0778) | 3 (4.55%) |
| | | | A4_property | 0.031 | 0.153 (0.0306) | 0.241 (0.0897) | 0.163 (0.0250) | 15 (22.7%) |
| | | | A5_investigation | 0.04 | 0.330 (0.0849) | 0.220 (0.738) | 0.168 (0.0929) | 15 (22.7%) |
| | | Total Visit Duration | A2_nations | 0.049 | 0.510 (-) | 0.208 (0.0699) | 0.185 (0.00707) | 7 (10.6%) |
| | | | A3_allocation | 0.051 | - (-) | 0.220 (-) | 0.295 (0.0778) | 3 (4.55%) |

| Test item | AOI | Eye- tracking Variable | Content word | Omnibus significance (*p* value) | μ 0 (SD) | | μ 1 (SD) | No. of included cases |
|---|---|---|---|---|---|---|---|---|
| | | | A5_transportation | 0.058 | 0.258 (0.0922) | 0.451 (0.303) | 0.637 (0.197) | 14 (21.2%) |
| | | Visit Count | A1_generations | 0.086 | 2.33 (1.16) | 1.67 (1.07) | 3.67 (2.94) | 21 (31.8%) |
| | | | A2_nations | 0.057 | 2.00 (-) | 1.00 (0.00) | 1.00 (0.00) | 7 (10.6%) |
| | | | A5_transportation | 0.06 | 1.00 (0.00) | 1.57 (1.13) | 2.33 (0.577) | 14 (21.2%) |
| | Verbs | Fixation Duration | Q_choose | 0.016 | 0.170 (-) | 0.195 (0.00707) | 0.150 (-) | 4 (6.06%) |
| | | Total Fixation Duration | Q_choose | 0.016 | 0.170 (-) | 0.295 (0.148) | 0.150 (-) | 4 (6.06%) |
| | | Visit Duration | Q_choose | 0.016 | 0.170 (-) | 0.305 (0.163) | 0.150 (-) | 4 (6.06%) |
| | | Total Visit Duration | Q_choose | 0.016 | 0.170 (-) | 0.305 (0.163) | 0.150 (-) | 4 (6.06%) |
| | Adjectives | Fixation Duration | A4_current | 0.078 | 0.230 (0.0829) | 0.283 (0.117) | 0.153 (0.0550) | 12 (18.2%) |
| | | | A5_public | 0.071 | 0.170 (0.0283) | 0.225 (0.0603) | 0.145 (0.0495) | 8 (12.1%) |
| | | Total Fixation Duration | A4_current | 0.046 | 0.755 (0.923) | 0.410 (0.173) | 0.175 (0.0436) | 12 (18.2%) |
| | | | A5_public | 0.071 | 0.170 (0.0283) | 0.225 (0.0603) | 0.145 (0.0495) | 8 (12.1%) |
| | | Visit Duration | A4_current | 0.094 | 0.293 (0.161) | 0.283 (0.117) | 0.153 (0.0550) | 12 (18.2%) |
| | | | A5_public | 0.071 | 0.170 (0.0283) | 0.225 (0.0603) | 0.145 (0.0495) | 8 (12.1%) |
| | | Total Visit Duration | A4_current | 0.046 | 0.758 (0.922) | 0.410 (0.173) | 0.175 (0.0436) | 12 (18.2%) |
| | | | A5_public | 0.071 | 0.170 (0.0283) | 0.225 (0.0603) | 0.145 (0.0495) | 8 (12.1%) |

*Note.* Cells with p values approaching significance are highlighted grey.