تأثير استخدام تقنية تتبع العين والتفكير بصوت عال لمساعدة متعلمي اللغة الأجنبية للتعرف الكلمات الانجليزية وفهم معناها في قراءة النصوص

جمعت الدراسة الحالية اختبارات التفوق والتكافؤ التقليدية للتحقيق في تأثير استخدام تقنية تتبع العين والتفكير بصوت عال لمساعدة المتعلمين على معرفة وفهم معنى الكلمات الإنجليزية في قراءة النصوص. استخدمت هذه الدراسة اختبارات مكافئة، والتي من المفترض أن تكون أكثر ملاءمة لإشراك المتعلمين غير المتفاعلين. تم اختيار مستويات عالية من متعلمي اللغة الإنجليزية لقراءة النصوص القصيرة التي تحتوي على كلمات زائفة. قسم المشاركين إلى ثلاث مجموعات: مجموعتان تجريبية (أي تتبع العين والتفكير بصوت عال)، ومجموعة مراقبة واحدة. وكشفت النتائج أن أيا من المجموعتين التجريبية لم تظهر أداءا قويا عند فهم معنى الكلمات الإنجليزية والتعرف عليها. ومع ذلك، كان التفكير بصوت عال أفضل إلى حد ما على النقيض من تتبع العين.

The Effect of Using Eye Tracking and Thinking: A loud technique to assist foreign language learners to recognize and understand the meaning of English words in reading texts

Dr. Elham Sweilam Ahmad Desouky

October 6 University, Faculty of Education

Curriculum & Instruction (TEFL) Department

Email: elesweilam@gmail.com

Abstract

The current study combined traditional superiority and equivalence tests to investigate the effect of using eye tracking and thinking, a loud technique for helping learners recognize and understand the meaning of English words in reading texts. This study used equivalent tests, which are supposed to be more suitable for engaging unreactive learners. High levels of English learners were selected to read short texts enclosing pseudo-words. The participants were divided into three groups: two experimental groups (i.e., eye tracking and thinking aloud), and one control group. The findings revealed that neither of the two experimental groups showed a strong performance when recognizing and understanding the meaning of English words. However, thinking aloud had a somewhat better impact in contrast to eye tracking.

KEY WORDS: Foreign language learners; English words; eye tracking; thinking aloud

The Effect of Using Eye Tracking and Thinking: A loud technique to assist foreign language learners to recognize and understand the meaning of English words in reading texts

Dr. Elham Sweilam Ahmad Desouky

October 6 University, Faculty of Education

Curriculum & Instruction (TEFL) Department

Email: elesweilam@gmail.com

Jegerski and VanPatten (2014) claimed that applied linguists should be able to study various cognitive processes involved in language learning and use appropriate advanced technology. Therefore, the current study aimed to develop a better comprehension of the exact information given through different methods of research (methodologies) as well as the drawbacks of each. The main target of the present research was to explore whether using methods of thinking aloud and eye tracking would have a positive effect on foreign language learners. To this end, the study used both traditional superiority tests and equivalence tests, based on the idea that equivalence tests would be beneficial for two reasons: to engage unreactive learners and to identify the non-reactivity of the research methodology used.

Review of Previous Research

The researcher divided the review of previous research into three sections: studies on thinking aloud with eye tracking, studies on eye tracking as a method of study, and research on the thinking aloud method.

Thinking Aloud with Eye Tracking

The thinking aloud strategy uses the oral method as the standard method of communication, by allowing the candidate to verbalize his or her ideas while performing a certain mission. The method is relatively new in foreign language acquisition (FLA) compared to its long-standing history in cognitive psychology (Bowles, 2010). On the other hand, eye-tracking strategies have been studied by recording candidates' eye movements as they see the information on a laptop screen (on most occasions).

Eye tracking has been an evolving method of study for years and is currently gaining much popularity in the FLA field (Godfroid & Schmidtke, 2013); in part due to it becoming much easier with the progression and abundance of the needed technology. An important reason why thinking aloud and eye tracking are common methodological companions is that both allow the researcher to gather complementary data (Leow, Grey, Marijuan, & Moorman, 2014); thus, they are commonly interlacing methods. Holmqvist et al. (2011) reported that thinking aloud is the most commonly used auxiliary data source in eye-tracking studies. A large number of former studies have found that an important value is added if the two measures are used in conjunction (Smith, 2012).

For example, in a reading-based study for native speakers in Finland, Kaakinen and Hyönä (2005) assessed reading purpose by requesting that children read about a rare disease that may have been contracted by their colleague. The researchers observed that the participants required longer first-pass reading times compared to the passage on another disease. The learners did not show a higher level of comprehension in their thinking-aloud strategies either. Despite that request, the related data did not show increasing evidence of better comprehension than request-unrelated ones. In this context, longer eye-fixing times acted as an intermediate factor between request-related data and the candidate's depth of comprehension as demonstrated by the thinking-aloud verbal reports.

Smith's (2012) study was designed for English learners in a foreign language class and observed learners' reactions during a computer-mediated chat. Nine pairs of learners were selected to give a written rundown of a clip to a native English speaker; at the same time, their eyes were tracked through a heat map analysis. In addition, their comments were followed during the text chat in response to a stimulated system designed for recall. Smith identified a direct correlation between the two measured factors (i.e., eye movement and recall stimulation).

Godfroid and Schmidtke (2013,p.196) studied the use of eye tracking and retrograde reporting in an assessment of attention and awareness as the chief elements of notice during the interpretation of incidentally placed novel vocabulary within an English reading class for non-native speakers. The learners' attention to new words was recorded in terms of eye-fixation time and the concluded level of awareness as reported by the student. The study concluded that eye-tracking interpretation and retrograde reporting were able (in a quite similar output) to predict a learner's cognition of novel words, as evidenced in a sudden post-test. The researchers figured out a very important point when they stated that verbal reporting alone was enough to predict word cognition; however, entering both data forms into a certain statistical module offered deeper comprehension of the study's objectives. The researchers also predicted that, as a result of the previously described information, dozens of studies using what has been described literally as "data triangulation between eye tracking and thinking -a loud method" would emerge in the near future, but more than a single recommendation should be considered, including:

- The timing of reporting "whether during or after the eye-tracking strategy" has to be accurately tailored;
- A simultaneous eye-tracking and verbal reporting combination has to be designed to be applicable if any support is planned to help learners' recall; and
- Eye tracking should not be performed (not accurate enough) when loud thinking is processed as the learners may move their heads during speech. Eye fixation may be prolonged due to the time the speaker takes to verbalize certain words; consequently, it will not accurately mirror the time taken to perform the main task as it reflects the verbalization of thoughts in addition to concurrent social and psychological issues, (p.196).

Thus, simultaneous thinking-aloud and eye-tracking methods are obscured by many obstacles. From this perspective, it is more applicable to show video or screen shots of a learner's eye as a signal of recall at the end of the given task; however, it remains unclear if the described visual aids would improve the quality of the received data or lead to thought fabrication.

In the current study, the feedback of thinking aloud and eye tracking using an inter-subject design was performed. Despite the fact that the use of eye-tracking methods is quite promising, minimal solid information is available regarding its effect on the process of reading. The same is true of the effect of thinking aloud compared to that of eye tracking. Yet the effects of the thinking-aloud method have been studied more frequently.

Eye Tracking as a Method of Study

Eye tracking as a method following the process of cognition is a well-known principle in the field of psychology, despite the fact that the researcher found that most studies on eye-tracking strategy in psychology have used theory building rather than study reactivity. More than one model has been manufactured to predict the sites as well as the times at which the reader will move his/her eyes when reading (Engbert & Kliegl, 2011). As a result of these models, a truly novel assumption has been built regarding how cognition affects eye movements; thus, eye tracking as a psycholinguistic tool of research has been assessed in terms of validity. The use of eye tracking as a study method relies on the assumption that eye movement and cognition are related to each other to some extent. This idea has been studied in cognitive-control models (Reichele et al., 2013) demonstrating that eye movements reflect either serial or parallel attention allocation during the process of reading.

Eye tracking as a method of study is gaining popularity as it is supposed to mirror cognition more than any other method can (Just & Carpenter, 1980). Moreover, it is easier than other techniques, where information collected through the Internet is used as we can enclose text stimuli, as in the moving window self-paced technique. In addition, it is much easier for the participant as he/she is not asked to perform additional tasks (e.g., pressing a button), which helps avoid bias and task interruption by the learners (Dussias, 2010). As a result, eye tracking has been plotted as the closest experimental technique to life reading (Van

Assche, Drieghe, Duyck, Welvaert, & Hartsuiker, 2011). On the other hand, participants using eye-tracking techniques reported difficulty more than once when wearing head-mounted watchers for prolonged periods as well as difficulty experienced from devices used to minimize head movements, which may add an unnatural feeling to the experience.

Godfroid and Spino (2015) preferred to tell the participants that they were participating in an experiment in order to pay attention and know that their eye movements were important for the experiment. When being monitored or recorded, people change their behavior. Those whose eye movements are being tracked may do the same by reading correctly and carefully when being observed. Therefore, the participants must be calibrated again and again during the experiment to obtain accurate and correct data. The mentioned elements could theoretically change the way that the participants read the text, which may affect their understanding of the texts.

Many researchers identified specific elements that could affect reading tasks when using eye tracking, such as the size of the font and line width. They found that spacing enhanced the speed of reading while improving the rapid movements of the eye. Spinner, Gass, and Behney (2013) designed two studies to investigate the effect of variety in font size, font type, and display model to determine whether the different ways of presenting the linguistic motive in an eye-tracking experiment would lead to different results. The results of the two studies showed the importance of the display model; in addition, the ideal font size was 24, as used in their second study.

Thinking Aloud

As the discussion thus far indicates, studies on eye tracking have been rare because it is a method following the process of cognition—a well-known principle in the field of psychology—which is different from methodological knowledge. On the contrary, numerous studies have concentrated on thinking aloud; this may be due to the nature of thinking aloud, which demands that participants be involved in another task at the same time. Eye tracking does not do this. Thus, the need for a secondary task stood behind the reactivity that appeared with thinking aloud (Morgan-Short, Heil, Botero-Moriarty, & Ebert, 2012). In contrast, Bowles (2010) refused the idea of reactivity with thinking aloud and claimed that it is not easy to decide whether reactivity appears with thinking aloud or not because it depends on the different variables examined in the study. To illustrate her point, she provided many elements that could affect reactivity, such as the type of task, type of language used in the task, type of verbal report, and proficiency. In her reactivity studies, she found that these elements affected four areas: understanding, time used on the task, the form of learning received, and the form of learning production.

All the research that investigated the reactivity of thinking aloud used traditional tests, such as *t*-test and ANOVAs. These traditional tests examine the null hypothesis and consider whether a difference exists between the thinking-aloud group and control group. As a result, researchers have refused the null hypothesis and approved the alternative to support reactivity. They used these tests to prove that, when they can disapprove the null hypothesis, the thinking-aloud group is reactive and, when the opposite happens, the thinking-aloud group runs nonreactive. Nevertheless, being unsuccessful in rejecting the null hypothesis does not give any evidence on the correctness of the hypothesis; it only gives support to reject the alternative. Furthermore, it is not theoretically accepted for considering thinking aloud as not being reactive just because the researcher was unsuccessful in rejecting the null hypothesis (Mascha & Sessler, 2011).

To overcome this problem, the current study used both tests—superiority and equivalence—to benefit from the merits of each one. The role of the equivalence tests is to discover whether the two groups are subject to the same treatments and conditions or not; at the same time, these tests are not suitable for emphasizing the reactivity of any methodology. Meanwhile, the superiority tests can be used to prove the reactivity of certain methodology, but unfortunately these tests are not sufficient for proving non-reactivity. Therefore, it was important to combine the two tests (Godfroid & Spino, 2015). When dealing with superiority tests, Larson-Hall,(2010) urged researchers to show the effect sizes because they may affect the results. The effect sizes can also play an important role in assisting readers in understanding the statistically significant impacts and appreciating their importance in their practical lives.

Yet Morgan-Short et al. (2012) clarified that effect sizes can be effective only when the sample is sizable and the statistical power is high enough that the differences between the groups can be significant. The researchers' opinion was based on their study in which they sought to determine whether learners are able to pay attention to the form and meaning of words at the same time when reading small texts. The results of their study showed that the effect size of the think-aloud sets were very small even though the scores of these groups were statistically high on comprehension tests. It can be concluded that effect sizes rely on superiority tests; therefore, both of them have the same target of showing non-reactivity. The effect sizes furnish important information, but it is completely descriptive, so the researchers cannot use it instead of equivalence tests. Despite this, it is important to display the effect sizes in superiority tests regardless of whether the results were significant or not. In addition, when the difference between two groups in the experiment is not significant, the best solution is to invert the two hypotheses and apply the equivalence tests (Plonsky, 2013).

Present Research

The present research was designed to evaluate the performance of the participants when recognizing and understanding English words while reading texts after applying eye-tracking and thinking-aloud techniques. Grabe & Stoller, (2011) mentioned that recognition and comprehension are not at the same level of knowledge because comprehension is considered a higher level than recognition; hence, the methodology of this research could be interactive with only one approach. To achieve the purpose of the study, the researcher developed several research questions to help ensure that this study is more accurate. These questions were formulated in two areas, as follows.

Word Recognition

- Is there a significant difference among thinking-aloud, eye-tracking, and control groups?
- Is there a significant statistical equivalent manner among thinking-aloud, eye-tracking, and control groups?

Word Comprehension

- Is there a significant difference among thinking-aloud, eye-tracking, and control groups?
- Is there a significant statistical equivalent manner among thinking-aloud, eye-tracking, and control groups?

This study used superiority tests to answer the questions related to the significant differences between the groups to measure reactivity whereas the other two questions were concerned about non-reactivity, so equivalence tests were used. This study tried to benefit from the results of other studies in two ways. As eye tracking does not require participants to engage in a subsidiary task (Dussias, 2010), this study hypothesized that, when dealing with the comprehension of the text and word recognition, eye tracking would be nonreactive. In addition, this study hypothesized that, when dealing with comprehension of the text, thinking aloud would be nonreactive, but the reaction could be positive on the posttest in terms of recognizing the forms of the lexical goal (Bowles, 2010).

Method

Participants

The learners who participated in this study were 51 males and females who ranged in age between 18 and 18.5 years old. All were in their first year of college in the English department of a faculty of education. All were Arabic speakers who had studied formal English for at least seven years. The researcher chose participants randomly, without any plan or biases, and divided them into three groups: 14 in the eye-tracking group, 14 in the think-aloud group, and 23 in the control group.

Materials

The researcher divided the materials used in this study into three parts: materials used in the reading task, content for the posttest, and word recognition for the posttest. These parts are discussed next.

Reading task

The present research used a leisurely style when the learners were asked to read 20 short, unconnected texts taken from English newspapers. Only 12 short texts were used for experimental items, and they contained 12 goal words, one word in each text (i.e., nine new words and three words known to the participants). Each learner read nine texts with new word once in addition to the other three control texts with target words. To assist the participants, the researcher broke down the new words and gave synonyms for six of them after expanding each text to three pages. The target word was always in the middle of each text on the second page (see Appendix A).

Content of the posttest

The study used 20 texts; one sentence was chosen randomly from every text for use in the posttest, so the posttest contained 20 different sentences. The participants answered the posttest by choosing one answer from three options given on the posttest (i.e., true, false, and I don't know). The researcher gave the participants some instructions before they responded to the posttest. First, participants were asked to answer the posttest in order, without any modification; second, they were asked to read every sentence carefully, without hurrying, and then determine the correct answer for each sentence.

The present research tried to determine whether the study questions measured the knowledge that participants obtained by reading the short texts or whether any learner could correctly answer the study questions through world information. Therefore, the researcher chose some other new learners (N = 41) who share the same features as those who participated in the study. These outsider participants were asked to answer the questions directly without reading the texts. The results showed that they correctly answered 17.81% of the 20 sentences (SD = 5.31, item range =2.47–46.3), and mean response accuracy was 14.085% (SD = 6.805, item range = 2.47–27.16). The results of the outsider participants showed that they were not able to predict the right answer of the comprehension questions without reading the texts. Furthermore, the results showed the validity of the reading comprehension measure used in the current study.

Word recognition of the posttest

The current study designed a word recognition posttest to identify to what extent the learners were able to recognize the nine experimental words and the three control words in the 12 texts they read before. At the beginning, the researcher showed a sentence from the text that the participants had encountered before; the target word included in this text but had been replaced with a space. The researcher gave the learners a list of words and asked them to choose the correct target word from this list. They had 30 seconds to read the sentence and determine the missing target word. The ability to guess correctly was limited due to the high number of distractors. The distractor items were divided into two sections: old words participants had encountered in the texts before (M = 3.22, range = 2–4.5) and new words encountered for the first time (M = 4.32, range = 3–5). To limit the difference between the experimental test items and control test items, a few existing words were included in most of the distractor lists for the experimental items (M = 0.96, range = 0–1.5). The researcher randomly distributed the three types of words (old/new/existing) among the three lists, and the content of each was different. The participants were not able to select the right word because simple recognition of the old words in the response options given was not enough to help select the right choice. Therefore, the researcher changed the approach by asking the participants to match their selected word and the word meaning in the original context.

Procedures

In this study, the researcher used a quasi-experimental design by dividing participants into three groups: eye-tracking group, thinking-aloud group, and control group. The three groups began by reading the texts to the end; the researcher then surprised them by administering two posttests—one for word recognition and the other for content—without indicating the records of thinking aloud or eye tracking. Finally, they were asked to complete a language history questionnaire. The researcher collected data from the first two groups by meeting them individually; data were collected from the control group through two whole-group sessions. The conditions of the three groups concerning type of font (Century Gothic), size of font (14 point), and spacing (double line) were the same. Both experimental groups read the short texts, but in different conditions as the eye-tracking group read the texts on a computer screen while the movements of the eye were registered using Eye Link II attached to the participants' heads.

The participants in the thinking-aloud group were asked to express their thoughts by speaking out loud. In dealing with the thinking-aloud group, this study used Type 1 (non-metalinguistic) verbalizations (Ericsson, 1993), and participants were not required to give any clarifications or good reasons about their thoughts. Participants in this group were also directed to use their first language (i.e., Arabic) or a mix of Arabic and the foreign language English. They read the short texts without recording their eye movements. The researcher gave each participant in the thinking-aloud group a flower and asked participants to raise the flower to give the researcher a chance to determine the part of the reading according to their spoken thoughts. The participants were asked to continue thinking out loud even when their face showed that they were thinking about something but did not want to say out loud. In addition, the researcher gave participants in the thinking-aloud group two texts as a model for being trained on thinking out loud.

All groups read one short text as a warm-up at the beginning. Unlike the two experimental groups, the control group remained completely silent without using thinking aloud or eye tracking.

Data Analysis

As previously discussed, the current study used superiority and equivalence tests together. The role of the superiority tests was to answer the questions related to the significant differences between the groups to measure reactivity. The equivalence tests were used to identify the equivalence margin, whereas superiority tests were used as a starting point to evaluate the performance of the participants in the control group. The role of the equivalence tests was different as they are conceptually appropriate for demonstrating nonreactivity. The current study clarified the analysis of the two tests as follows:

Superiority tests

The superiority tests used AGEE logistic regression (Hardin & Hilbe, 2013) to test twofold dependent variables (giving scores for the word recognition and text comprehension items). The AGGE logistic regression differs from the other tests because it works on the level of the item, making the tests more powerfully built because the variance that relies on the item can be calculated (see Appendix B).

One advantage of the AGGE is that it takes into consideration all the non-independent observations. The current study tried to measure the two prime effects by putting the data collection method in the same row with test item and then combining them with response accuracy. In this context, the control group was used as a starting point case. Odds ratios (Ferguson, 2009) were used to count the effect size to gauge eye tracking and thinking aloud, and then record the differences between them. In addition, odds ratios were used to calculate the changes that occurred in the accuracy of the control group based on the accuracy test for certain data collection methods. Ferguson (2009) identified the ranks of odds ratio: no effect = 1 whereas more than 3 or less than 0.33 were considered to have a powerful impact to some extent. According to Field (2013), an identical data collection method will be reactive when the confidence interval reaches 95% while the odds ratio does not stretch to 1. Based on this, if the odds ratio is more than 1, the data collection method will be nonreactive; when the odds ratio is less than 1, the data collection method will be positively reactive.

Equivalence tests

In the equivalence tests, it is important to establish the acceptance standard in a way that makes any treatment between equivalence and acceptance zone leave the unmentioned empirical effect on the findings while the meaningful difference of these treatments falls outside this time (Godfroid & Spino, 2015). Brown (2005) claimed that it is suitable for the equivalence tests to use the standard error of measurement (SEM) as an acceptance standard for two reasons. First, it is used as an equivalence standard for all test participants to examine the measurement error of the test. Second, if the test was given to the same learner several times, the learner's score may fall. Thus, the role of the SEM is to make a range around the learner's score. Like an effect size in the superiority tests, there is an equivalence band in the equivalence tests: If the null hypothesis is refused, the evidence for practical equivalence is strong and the equivalence band narrow. The present research adopted a stricter criterion for equivalence by diminishing the SEM; in addition, the items of the test were trimmed to ensure trustworthiness of the items. Cronbach (1951) and Guttman (1945) identified the reliability for the word items in their studies. For the 11 items, the reliability was 0.69; for the 12 word items, it was 0.56. The current study analyzed whole data sets and minimized ones in the superiority tests; however, the analysis of the two data sets did not show any difference in results.

Superiority tests versus equivalence tests

The equivalence tests and superiority tests valued the differences between the groups and the differences of confidence intervals as well. The hypothetical mean scores for the experimental groups combined with confidence intervals are a = 0.05; the superiority tests depended on 95% confidence intervals while the equivalence tests relied on 90% confidence intervals (Godfroid & Spino, 2015). For the experimental group, the superiority test was considered reactive when it did not spread to the starting point of the control group and the mean score was at the 95% confidence interval. Concerning the equivalence tests, the values of the point was sufficient to show whether the two experimental samples in a study were equivalent (falling within the equivalence band) or not.

Results

Word Comprehension

The current study posed two research questions concerning word comprehension. The superiority and equivalence test results were analyzed to answer these questions.

Superiority tests

As previously discussed, the participants answered the posttest by selecting from three choices (i.e., yes, no, or I don’t know). The scores were divided as follows: one mark for the right response and no marks for the false response or I don’t know choice. The results revealed that the average correct response for the participants was 9 out of 11 sentences (SD = 2.69), which reflected a good comprehension of the text. The best comprehension of the text was in the thinking-aloud group (M = 4.395, SD = 0.915), followed by the control group (M = 4.175, SD = 1.13) and eye-tracking group (M = 4, SD = 0955). The differences between scores were small to some extent: Wald $\chi^2(2) = 1,475$, $p = .23$. There were no significant differences between the comprehension of the text and the task of data collection technique after combining generalized estimating equations (GEE) and test items to use them as repeated-measures variables. Data collection method and test items were combined for use as predictors.

The odds ratios for the three groups were different as they diminished in the eye-tracking group to reach 0.185, 95% CI = (0.235, 0.57), *p* = .17, compared to the thinking-aloud group's 0.52, 95% CI = (0.325, 0.845), *p* = .86, which means it increased. The participants' performance in the control group reached 80%, eye tracking 74%, and thinking-aloud 81%. Based on this, no significant differences emerged among the three groups concerning participants' performance as none of them outperformed others; hence, compared to the control group's results, neither eye-tracking nor thinking-aloud groups had any significant effect on the comprehension of reading text (see Appendix B).

Equivalence tests

The present research used the same band of equivalence values (0.70, 0.89) as Godfroid and Spino (2015) to act as equivalent values to the performance of the control group. These values depend on the SEM for test comprehension and GEE regression coefficient for the performance of the control group (see Appendix C). The results showed that the three groups were functionally equivalent in text comprehension, as their estimated values fell between the equivalence bands. As a result of the stretched confidence intervals to the equivalence band, the effect was not reliable at the public standard. The inferential statistics for equivalence tests were used to complete the confidence interval in both parts—namely, the lower boundary and upper boundary (Mascha & Sessler, 2011). The lower boundary for the eye-tracking group was Dunnet's *t* (1.495) = 0.165, *p* = .46 and upper boundary Dunnet's *t* (1.495) = -0.715, p = .13. The upper boundary for the thinking-aloud group was Dunnet's *t* (1.495) = -0.43, *p* = .29 and lower boundary Dunnet's *t* (1.495) = 0.505, *p* = .25.

Word Recognition

This section discusses the two previous research questions concerning word recognition. The results will be used to answer these questions in terms of the superiority tests and the equivalence tests.

Superiority tests

In the word recognition posttest, the learners earned one mark for the right response and no marks for the other answers. The test was considered somewhat difficult as 930 observations showed 270 (29%) right answers and 671 (72.1%) wrong answers. The average recognition for the participants was 1.265 (SD = 0.9) out of 9 pseudo-word items. More specifically, the word recognition in the thinking-aloud group was the best (M = 1.57, SD = 1), the control group ranked second (M = 1.23, SD = 0.9), and the eye-tracking group (M = 1.02, SD = 0.715) ranked last. Based on Wald χ^2 (2) = 5.925, *p* = .003, it was clear that the effect of the data collection technique proved to be significant after combining GEE and test items to use them as repeated-measures variables as well as combing the data collection method and test items to use them as predictors. The odds ratios for the two groups were different compared to the control group; they diminished in the eye-tracking group to reach 0.345, 95% CI = (0.21, 0.575), *p* = .16, while in the thinking-aloud group it reached 0.855, 95% CI = (0.53, 1.385), *p* = 0.03, meaning it increased. The participants' performance in the control group reached 25%, compared to 19% in the eye-tracking and 37% in the thinking-aloud groups. However, the thinking-aloud group showed a small positive impact that was statistically significant, although the Kruskal-Wallis test was not able to reveal the reactive impacts of this technique (see Appendix B).

Equivalence tests

The present research used the same band of equivalence values (.21 and .27) as Godfroid and Spino (2015) as equivalent values for the performance of the control group. These values depend on the SEM for test comprehension and GEE regression coefficient for the performance of the control group (see Appendix C). For the word recognition, compared to the control group, both eye-tracking and thinking-aloud groups were functionally reactive as their estimated values fell outside the equivalence bands. The eye-tracking group seemed to do worse on word recognition as the lower boundary was Dunnet's *t* (1.495) = 0.15, *p* = 0.53, and the upper boundary was Dunnet's *t* (1.495) = -0.485, *p* = 0.26. The upper boundary for the thinking-aloud group in word recognition was Dunnet's *t* (1.495) = -0.36, *p* = 0 .66; the lower boundary was Dunnet's *t* (1.495) = 0.655, *p* = 0.16. Thus, this group seemed to do better in word recognition.

Discussion

The present research combined two tests, superiority and equivalence tests, to provide an effective conceptual analysis about reactivity and nonreactivity in the studied groups. The current study sought to determine whether any efficacy emerged in foreign language learners' ability to recognize and comprehend English words while reading while using eye tracking and thinking aloud. Concerning word recognition, the results revealed that the thinking-aloud group was positively reactive with a small effect size. Although the GEE analysis suggested that the eye-tracking group was nonreactive, the equivalence test considered it to have negative reactivity because the participants in this group recognized fewer words than those in the control group.

Eye-Tracking Reactivity

The results of the eye-tracking group showed that recording eye movements did not hinder or facilitate the capacity of the foreign language learners when dealing with short texts, although the reading situations in this group were uncommon because learners came to the computer lab to read and were asked to stop for calibration several times. The calibration in this study required standing for nearly 20–25 seconds each time when reading the text, and participants had to recalibrate 12 times for each text. Notwithstanding these conditions, the participants' performance in this group (compared to the control group, who had different conations as the learners read on paper in their classroom without obstructions) was the same as the control group in terms of reading comprehension.

Concerning word recognition, both superiority and equivalence tests kept the null hypothesis because none of them gave a firm decision about the reactivity or nonreactivity of eye tracking. The equivalence test suggested negative reactivity while the GEE analysis suggested nonreactivity. The results of the current study showed that any effect of eye tracking will be smaller than the effect of the thinking-aloud technique. In word comprehension, neither technique was reactive, despite using double-spaced 14-point font for reading texts for all groups. This design was used for eye tracking to guarantee measurement accuracy; however, this study used it with all groups to be able to compare the three groups. The model in this study may have simplified the process of reading as this font and double-spaced text enlarges the font size more than usual. The participants in the current study were mindful of the importance of their eye movements for the study; hence, they may have tried to appear in a good form in front of the researcher by reading fluently. This probability makes the researcher think that the small, negative effect on word recognition in the eye-tracking group may be due to this reason.

Thinking-Aloud Reactivity

The thinking-aloud group was reactive in word recognition only, but not on comprehension according to the superiority and equivalence tests. Godfroid and Spino's (2015) results were in line with the current study as they found that thinking aloud has no effect on comprehension. To determine which thinking-aloud elements help achieve word recognition, the researchers collected all the recorded verbal protocols for the thinking-aloud group, and then categorized them into two parts: awareness versus pronunciation. The results showed that the best design was the one with both pronunciation, Wald χ^2 (1) = 6.53, $p <$ 0.001, and awareness, Wald χ^2 (1) = 1.905, p = 0.051; test items also predicted word recognition, Wald χ^2 (11) = 36.895, p < 0.001. Thus, word recognition had a moderately strong influence by using pronunciation while thinking aloud: odds ratio =1.445, 95% confidence intervals = (0.815, 2.565), p < 0.001. The word recognition was independent from any significant effect of awareness, odds ratio = 1.04, 95% CI = (0.5, 2.17), p = 0.051, which means that word recognition was enhanced by using pronunciation and awareness.

The results of the current study showed that the thinking-aloud group differed from the other two groups who read silently because carrying letter-sound conversations all the way to the articulation level helped the thinking-aloud group do better in word recognition while also developing memory quality; therefore, verbalizing one's ideas or opinions may enhance memory and the encoding of language forms while thinking aloud. In Bowles's (2010) study, the results revealed that benefits of verbalization may spread to morphosyntactic target forms as well. Unlike the other groups, thinking aloud was the only group where the participants uttered some of the target words clearly when reading the texts; thus, for word recognition, pronunciation was a medium-strong predictor. Coltheart, Rastle, Perry, Langdon, and Ziegler (2001) claimed that, during silent reading, phonological information is activated, but this activation still is invisible.

Equivalence Tests

Equivalence tests are usually used with large samples, so the results from the equivalence tests may be inaccurate due to the small sample size used in the present study. To make the results of this study statistically trustworthy, the sample needs to be larger (Streiner, 2003). The present research was not successful in proving any significant differences with wither equivalence or superiority test. However, equivalence tests can give conceptually sound support for the matching of participants on two sides of the linguistic properties of the stimuli (i.e., length, frequency, phonotactic probability, predictability) and biographical variables (i.e., amount of exposure, age, proficiency level) (Godfroid & Spino, 2015). This study aimed to show that the equivalence tests can have a place in foreign language acquisition not only in reactivity research, but also in any research that tries to demonstrate the equivalence of two treatments (e.g., types of feedback) on the measure of the results. Wellek (2010) suggested using a noninferiority test (the one-sided counterpart of an equivalence test) to prove that a novel and perhaps more economical intervention (e.g., online training) is statistically noninferior to a traditional approach.

Conclusion

The present research aimed to introduce eye tracking within reactivity research and answer the reactivity question using appropriate statistical tests. The results revealed that, in terms of word comprehension, neither eye-tracking nor thinking aloud were reactive. In word recognition, the situation was different, as the thinking-aloud group was positively reactive with a small effect size. The results of the eye-tracking group could be divided into two parts; GEE showed the eye-tracking group to be nonreactive while the equivalence test suggested it showed negative reactivity. This area of research needs more investigations to find out whether eye tracking has any subsequent effects on participants' performance when reading.

References

Bowles, M. A. (2010). *The think-aloud controversy in second language research.* New York: Routledge.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* New York: McGraw-Hill.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108,* 204–256.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297- 334.

Dussias, P. E. (2010). Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics, 30,* 149–166.

Engbert, R., & Kliegl, R. (2011). Parallel graded attention models of reading. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 787–800). New York: Oxford University Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Ferguson, C. J. (2009). An effect size primer: A guide for clinians and researchers. *Professional Psychology: Research and Practice, 40,* 532–538.

Field, A. (2013). *Discovering statistics using SPSS* (4th ed.). Thousand Oaks, CA: SAGE.

Godfroid, A., & Schmidtke, J. (2013). What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports and vocabulary learning scores.

Honolulu: University of Hawai'i at Manoa, National Foreign Language Resource Center.

Godfroid, A., & Spino, L. (2015). Reconceptualizing reactivity of think-alouds and eye-tracking: Absence of evidence is not evidence of absence. *Language Learning, 65*, 896–928.

Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading* (2nd ed.). Harlow, UK: Longman.

Guttman, L. (1945). A basis for analysing test-retest reliability. *Psychometrika, 10,* 255-282.

Hardin, J. W., & Hilbe, J. M. (2013). *Generalized estimating equations* (2nd ed.). Boca Raton, FL: CRC Press.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye-tracking: A comprehensive guide to methods and measures.* Oxford, UK: Oxford University Press.

Jegerski, J., & VanPatten, B. (2014). *Research methods for second language psycholinguistics.* New York: Routledge.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 326–354.

Kaakinen, J. K., & Hyönä, J. (2005). Perspective effects on expository test comprehension: Evidence from think-aloud protocols, eye-tracking, and recall. *Discourse Processes, 40*, 239–257.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS.* New York: Routledge.

Leow, R. P., Grey, S., Marijuan, S., & Moorman, C. (2014). Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview. *Second Language Research, 30*, 111–127.

Mascha, E. J., & Sessler, D. I. (2011). Equivalence and noninferiority testing in regression models and repeated-measures designs. *Anesthesia & Analgesia*, *112*, 678–687.

Morgan-Short, K., Heil, J., Botero-Moriarty, A., & Ebert, S. (2012). Allocation of attention to second language form and meaning: Issues of think-alouds and depth of processing. *Studies in Second Language Acquisition, 34*, 659–685.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, **35**, 655–687.

Reichle, E. D., Liversedge, S. P., Drieghe, D., Blythe, H. I., Joseph, H., White, S. J., & Rayner, K. (2013). Using E-Z Reader to examine the concurrent development of eye-movement control and reading skill. *Developmental Review*, *33*,110–149.

Smith, B. (2012). Eye tracking as a measure of noticing: A study of explicit recasts in SCMC. *Language Learning & Technology*, *16*, 53–81.

Spinner, P., Gass, S. M., & Behney, J. (2013). Ecological validity in eye-tracking: An empirical study. *Studies in Second Language Acquisition*, *35*, 389–415.

Streiner, D. L. (2003). Unicorns *do* exist: A tutorial on “proving” the null hypothesis. *Canadian Journal of Psychiatry*, *48*, 756–761.

Van Assche, E., Drieghe, D., Duyck, W., Welvaert, M., & Hartsuiker, R. J. (2011). The influence of semantic constraints on bilingual word recognition during sentence reading. *Journal of Memory and Language*, *64*, 88–107.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and non-inferiority* (2nd ed.). Boca Raton, FL: CRC Press.

Appendix A
Sample Reading Text

People build houses as shelters to protect themselves from the weather. The materials used in building and the design planned differ according to different factors, mainly CLIMATIC/GEOGRAPHICAL/GEOGRAPHICAL OR CLIMATIC/CLIMATIC OR GEOGRAPHICAL conditions. Well-insulated buildings are adopted in regions with hot and cold seasons. These keep out both heat and cold. In rainy and snowy regions, the roofs slope steeply to allow the rain and snow to fall off easily. In regions of frequent earthquakes, such as Japan, buildings are made of wood and are specially constructed. Technology plays an important part in the design of buildings. It provides us with many facilities, such as means of communication. A mobile connects to the home telephone and to the microphone by the doorbell. Everything can be conducted though high-tech and modern equipment.

Comprehension Posttest right/wrong sentence: The materials used for buildings are similar.
Right answer: Wrong.

Note: The target zone appeared in regular print and has been capitalized. The learners read only one version of every text.

Appendix B
Kruskal-Wallis Tests

The current study has three groups with unequal sample sizes, and there were deviations from normality in the data; thus, the non-parametric test was used. When the results were not significant, this study used the following form to compute effect sizes for pairwise comparison:
Z (Field, 2013).

rExp. Group- Silent Control = $\overline{\sqrt{\text{nExpGroup + nSilentControl}}}$

These analyses were presented to give the readers the chance to evaluate the added value of using GEE logistic regression and equivalence tests.

Data of Comprehension

The comprehension results showed no difference between using GEE analysis and Kruskal-Wallis test as both revealed no significant differences among the three groups in comprehension: H (2) = 1.265, p = 0.28. Eye tracking had a small, negative effect (r = -0.06), while thinking aloud had a negligible, positive effect on text comprehension (r = 0.04)

Word Recognition Data

The Kruskal-Wallis test was not successful in discovering any significant differences among the three groups in word recognition, unlike the GEE analysis: H (2) = 2.075, p = 0.13. Eye tracking had a small, negative effect on word recognition (r = -0.055) while thinking aloud had a small, positive effect (r = .075).

Appendix C
Calculating the Equivalence Districts

The present research used three procedures to calculate the equivalence zone for the content and word test scores. The first procedure calculated the standard error of measurement (SEM), followed by accounting for the acceptance criterion and, finally, computing the equivalence zone. The following numbers are rounded to two decimals, which may calculate the differences of the small numbers different from the original computations.

Data of Comprehension

The mean score observed for the silent controls was 4.175 (*SD* = 1.13), with Cronbach's α = .69.

Standard Error of Measurement

SEM = $\sqrt{1\text{-r}}$

= 1.13 $\sqrt{1\text{-}}$ .69

= 0.63

To make the SEM like the GEE coefficients on the same scale; it was turned into a ratio (see Godfroid et al., 2015).

SEM in proportion = 0.63 ÷5.5 = .11

Value of Boundary =

SEM = .11

P (correct) control = 0.395

= 0.045

Equivalence zone centered on the control group's mean comprehension score

=

(0.395 – .045, 0.395 + .0.045) = (0.35, 0.44)

Word Recognition Data

The mean score observed for the silent controls was 1.23 (*SD* = 0.9), with Guttman split-half reliability = 0.56. Standard error of measurement

SEM = $\sqrt{1\text{-r}}$

= 0.9 $\sqrt{1\text{-}}$ .56

= 0.595

SEM in proportion = 0.595 ÷4.5 = .13

Value of Boundary =

SEM = .13

P (correct) control = 0.12

= 0.015

Equivalence zone centered on the control group’s mean comprehension score

=

(0.12 – 0.015, 0.12 + 0.015) = (0.105, 0.135)