Journal of Audiovisual Translation

Methodological Preparation of a Within-Subject Audiovisual Cognition, Reception and Perception Study

Senne M. Van Hoecke✉
University of Antwerp

Iris Schrijver✉
University of Antwerp

Isabelle S. Robert✉
University of Antwerp

Citation: Van Hoecke, S. M., Schrijver, I. & Robert, I. S. (2022). Methodological Preparation of a Within-Subject Audiovisual Cognition, Reception and Perception Study. *Journal of Audiovisual Translation,* *5*(1), 94–128. https://doi.org/10.47476/jat.v5i1.2022.163

Editor(s): J.L. Kruger

Received: September 14, 2020

Accepted: June 16, 2021

Published: April 19, 2022

Abstract

In the past decade, cognitive empirical AVT research has been on the rise. The majority of these studies are between-subject studies, focused on subtitles for the deaf and hard of hearing (SDH). The few experimental studies that are aimed at other audiences tend to have small sample sizes. Within-subject studies are rarely used in experimental AVT cognition, reception and perception research, although they can increase statistical power due to the repeated testing and shed light on the idiosyncratic nature of the matter. This paper pleads for the introduction of complementary within-subject designs by illustrating the contrasts between the within-subject and between-subject research design. Drawing from the broader spectrum of Translation Studies and the case of the Subtitles for Access to Education (S4AE) research project, this paper highlights obstacles in the preparation of a within-subject AVT cognition, reception and perception experiment and proposes a possible approach to prepare similar within-subject AVT studies.

Key words: audiovisual translation (AVT), research design, within-subject design, methodology, cognition, subtitle reception, subtitle perception.

✉ senne.vanhoecke@uantwerpen.be, https://orcid.org/0000-0003-0519-576X
✉ iris.schrijver@uantwerpen.be, https://orcid.org/0000-0001-6091-024X
✉ isabelle.robert@uantwerpen.be, https://orcid.org/0000-0002-8595-0691

1. Introduction

Audiovisual translation (AVT) has become a booming and multi-faceted research field over the past decades (Díaz-Cintas, 2020). The start of the new millennium saw the emergence of cognitive and empirical AVT studies, which tend to focus on subtitles for the deaf and hard of hearing (SDH) and audio description (Díaz-Cintas, 2020). Experimental research into the reception of AVT for other audiences and purposes other than language learning remains scarce (Díaz-Cintas, 2020; Díaz-Cintas & Szarkowska, 2020). However, as Díaz Cintas and Szarkowska (2020) point out, there is a need for such experimental research as it not only allows us to test new practices, but also enables us to verify old assumptions and theories. This research could "feed back straight into professional practices and processes" (Díaz-Cintas, 2020, p. 222). These scholars also underline the importance of sound methodologies, replicability and reproducibility in said research.

Adhering to the aforementioned importance of methodological transparency, replicability and reproducibility, the aim of this paper is to present the methodological preparation of a large-scale, within-subject (repeated measures) study into the reception and perception of and cognitive load posed by subtitles, the so-called Subtitles for Access to Education (S4AE) project. This article follows in the footsteps of a number of publications that lay out possible methods and methodologies or recommend certain approaches for experimental AVT reception research (e.g., Doherty & Kruger, 2018; Kruger et al., 2016; Kruger et al., 2015). Another important precursor is the position paper by Orero et al. (2018), which can be used as a solid guideline for research as it lists many previously conducted AVT studies, proposes numerous measurement tools and recommends various approaches and research designs. One design, however, receives relatively little attention in these publications, namely the within-subject design. What is more, within-subject designs appear to be scant in AVT cognition, reception and perception studies (for brevity purposes: AVTCRP studies) as a whole, with exceptions such as Jensema et al. (2000), Tsaousi (2016), Montero Perez (2019) and Liao et al. (2020). Slightly more frequent is the use of mixed designs, including both within-subject and between-subject components, e.g., Orrego-Carmona (2015), Gerber-Morón and Szarkowska (2018) and Szarkowska and Gerber-Morón (2018a, 2018b). These are, however, also limited in number. This article aims to shed light on the advantages and drawbacks of a within-subject design and the possible challenges that arise when preparing such a study.

This paper is structured as follows: Section 2 elucidates the contrast between within-subject and between-subject designs, based on literature sourced from the broader field of Translation Studies. In Section 3, the design of the S4AE project and the methodological preparations are explained in detail. The paper concludes with some methodological recommendations for future within-subject studies as well as a discussion of some limitations in our study.

2. Designs in Experimental AV Cognition, Reception and Perception Studies

The design of any experimental study is determined according to the main research question. Balling and Hvelplund (2015) classify three types of research design: (a) an independent (or between) groups design, comparing two groups; (b) a within-subject (repeated measures) design, examining the same group in various conditions; and (c) a functional relations design, focusing on relations between variables rather than participants' behaviour in various conditions. Combinations of these designs, mixed designs, are also possible. In this paper, we will mainly focus on the repeated measures design, contrasting its characteristics with the between-subject design. We chose this focus as we expect most readers to be familiar with between-subject designs, but not necessarily with within-subject designs, especially given the scarcity of such designs in experimental AVT research. For the basis of this paper, we draw from both research in AVT as well as from the broader field of Translation Studies.

A between-group (or between-subject) design is commonplace in AVTCRP studies. It tests different participants in various conditions or in one condition. There are numerous ways to plan a between-subject design, ranging from using a test group and control group in a regular and doctored condition (e.g., Bisson et al., 2014; Kruger & Steyn, 2014; Montero Perez, 2020; Szarkowska et al., 2011) to testing of participants by comparing conditions without control groups (e.g., Moreno & Mayer, 2002; Perego et al., 2010; Vulchanova et al., 2015). In contrast, a within-subject (or repeated measures) design is an experimental design in which the same participants are tested a number of times. Again, the specifics may vary depending on the research goal. Researchers can, for example, test the same participants in multiple conditions to examine how varying situations influence the participants (e.g., S4AE project, see Section 3.1) or they can compare before and after data in one condition (e.g., Montero Perez, 2019). All tests may take place in one session (e.g., pilot tests of the S4AE project) or may span over a longer period of time to assess developments (e.g., Moreno et al., 2011).

These designs have various contrasting advantages and disadvantages. The largest advantage of a within-subject design is the mitigation of variability due to the same participants being used for each condition (Mellinger & Hanson, 2017, p. 137). As a result of this lowered variability, the number of participants required to make reliable conclusions is smaller as well, which may be interesting for participant recruiting and possible recruiting costs as well. Between-subject designs are limited in their ability to account for differences between participants, which reduces statistical power in the case of smaller sample sizes (Mellinger & Hanson, 2017). Díaz-Cintas (2020, p. 7) stated that limited sample sizes are a present problem in the few experimental AVTCRP studies that are not focused on SDH. Complementary within-subject designs could therefore be a possible means to increase validity and reliability in experimental AVT research. Though the repeated testing might increase internal validity (i.e., accurate measurement), and reliability (i.e., experimental replicability and reproducibility) to some extent by repeatedly confirming certain findings, revealing patterns or showing consistency, it reduces external validity, i.e., ecological validity, as it is evidently conducted in a more experimental setting compared to a between-subject study (Frey et al., 1991; Saldanha & O'Brien, 2013, p. 33). The mitigation of personal variability can also be of benefit for the idiosyncratic

nature of particular research topics, such as perception and cognition, which we expect to be different for every individual. Within-subject designs could filter out any of these undesired individual influences and could, in combination with biographical surveys or participant profiling, also help identify influencing factors. In sum, within-subject designs would be a viable option to strengthen studies with smaller samples and mitigate, and possibly identify, influences resulting from personal differences. These two advantages have already been highlighted by Bernardini in 2001, when she addressed the frequent use of between-subject designs in TAP (Think-Aloud Protocols) based translation process research, often conducted with a very limited number of participants. Another advantage of a within-subject design is that it generally does not require control groups, which reduces the chance of contamination. Contamination occurs when an experimental group (un)intentionally passes on essential information about the experiment to the control group or vice-versa, which may mask the actual effects of what is tested. The reduced chance of contamination in within-subject designs can be considered a substantial advantage. It is nevertheless difficult to estimate how realistic and/or frequent this risk of data contamination is, since there have not been any reports – to our knowledge – in AVT research.

However, a within-subject design also has a number of drawbacks in contrast to a between-subject design. One contrast is the time required to adequately set up and execute an experiment. As Section 3.2 will reveal, it takes considerable effort to prepare a within-subject study compared to a between-subject study. The repeated testing also lengthens the experiment. Another contrast is that due to the extended length, a within-subject experiment becomes more prone to attrition and data loss (Mellinger & Hanson, 2017, pp. 7, 105). In the case of multiple tests at different points in time, participants may simply not be present for the repeated tests. Additionally, multiple tests increase the chances of data being unusable, especially in the case of eye tracking with poor calibration or low tracking ratios. A third drawback of the repeated testing is the influence of certain confounding variables (Charness et al., 2012). Mellinger and Hanson (2017, pp. 7, 105) distinguish three of these variables: (a) fatigue, (b) order effects, and (c) carryover effects. The multitude of tests can be tiresome for participants, which in turn may lead to decreased concentration and/or motivation, especially in later stages of the experiment. The participants' behaviour may also be different dependent on the order of the tests. Carryover effects imply that participants learn and improve over the course of an experiment, e.g., by conversing with one another, reading/watching relevant material (outside the experimental design) or becoming familiar with the way of testing, which may result in higher scores in later stages. Evidently, these confounding variables can significantly influence the results of a within-subject study, whereas they are less important in a between-subject study. One common solution is to employ counterbalancing. Nevertheless, Mellinger and Hanson (2018, p. 16) warn these confounding variables may still be present.

3. Project Preparation

3.1. Research Background, Goal and Design

To introduce the S4AE project, we would first like to illustrate its research background. Following modern globalization and migration, higher education institutions (HEIs) face increasingly multilingual and multicultural audiences. To cater to these audiences, many HEIs are starting to use English as a medium of instruction (EMI) (Wächter & Maiworm, 2014). The introduction of EMI, however, may have a negative impact on comprehension, cognitive load and retention for students less proficient in L2 English. Subtitles might help to overcome these language barriers and make EMI lectures more accessible. However, adding subtitles to the classroom implies that students suddenly must process a new source of visual information alongside the already present audiovisual information from the lecturer, the lecture slides, the whiteboard, etc. This increases the amount of information that needs to be processed and might thus be more cognitively demanding for students. Delving into this matter, the S4AE project builds on three previous studies exploring the effects of subtitles on comprehension and cognition in a standard educational context (Chan et al., 2019; Kruger et al., 2014; Kruger & Steyn, 2014) and aims to answer the following question:

> To what extent do the presence of subtitles (present/not present), the subtitle language (L1/L2), the level of L2 proficiency and the students' prior knowledge influence (1) the (perception of) cognitive load and (2) the comprehension and retention of an L2 English lecture?

To answer this question, the S4AE project can build on considerable between-subject AVT research[1] focusing on subtitle processing by the viewer (e.g., Bisson et al., 2014; Colm, 2008, 2009; de Linde & Kay, 1999; d'Ydewalle & De Bruycker, 2007; Kruger, 2013; Hefer, 2013a, 2013b; Kruger et al., 2018; Kruger et al., 2013; Moreno, 2017; Perego et al., 2018; Perego et al., 2010; Gerber-Morón et al., 2018; Perego et al., 2016a; Perego et al., 2016b), in which various approaches are used, e.g., eye tracking, electroencephalography, functional magnetic resonance imaging, self-report psychometric questionnaires, dual-tasks and recognition tests. Many of these studies also examine the effects of subtitles on comprehension and retention, which is also of interest when researching the use of subtitles in EMI classrooms. The existing research has predominantly focused on two aspects: (a) movie comprehension/retention (e.g., Bairstow, 2012; Bairstow & Lavaur, 2017; Birulés-Muntané & Soto-Faraco, 2016; Caffrey, 2008, 2009; Kruger, 2013; Lavaur & Bairstow, 2011; Szarkowska & Bogucka, 2019) and (b) comprehension/retention in a classroom context (e.g., Bianchi & Ciabattoni, 2008; Danan, 1992, 2004; Díaz-Cintas & Cruz, 2008; Montero Perez, 2020; Montero Perez et al., 2014; Moreno & Mayer, 2002; Vulchanova et al., 2015). These studies mainly examine language learning

[1] Due to space constraints, this body of research will only be briefly discussed in terms of approaches and measurement tools.

and as such mostly use vocabulary tests, language proficiency tests, word or scene recognition tests for retention and multiple-choice comprehension tests.

Interestingly, studies into subtitle processing and the effects of subtitles on comprehension and retention in an educational context that is aimed at content and not language learning seem scarce. We know only of the three studies mentioned earlier (Chan et al., 2019; Kruger et al., 2014; Kruger & Steyn, 2014). These use self-report effort, frustration and comprehension questionnaires, comprehension tests and eye tracking. They also distinguish visual attention from actual subtitle reading, using the Reading Index for Dynamic Texts (RIDT) developed by Kruger and Steyn (2014).

Complementing these three (between-subject) studies, the S4AE project revolves around a central within-subject design. However, following the advice of Mellinger and Hanson (2017, pp. 163–164), we extended the initial within-subject design to include between-group independent variables, which in turn allows us to assess the interactions between cognitive load and comprehension, and student L2 proficiency levels and prior knowledge of the subject as well. The inclusion of these variables does not alter the advantages, disadvantages or necessary preparation of a within-subject study that this paper discusses.

In this design, Dutch (Flemish) students will view three different recorded EMI lectures. These lectures will be provided in three conditions: (a) with intralingual (English) subtitles; (b) with interlingual (Dutch) subtitles; (c) with no subtitles. To minimise fatigue, order and carryover effects (Mellinger & Hanson, 2017, p. 105), the order of the lectures and the conditions will be counterbalanced completely. The students will watch the lectures individually in an eye tracking laboratory. Eye tracking will allow us to measure cognitive load and actual subtitle reading using Kruger and Steyn's (2014) RIDT as a complementary tool. After each lecture, the students will fill out an extended version of the psychometric questionnaire on cognitive load validated by Leppink and van den Heuvel (2015) and, subsequently, a comprehension test. Using both a psychometric self-report questionnaire and eye tracking to assess cognitive load allows triangulation of data from objective and subjective measures, as recommended by Orero et al. (2018). One month after the experiment, all participants will complete the same comprehension tests again to measure retention. The scores on the psychometric questionnaires and comprehension tests, as well as the eye tracking data, will be correlated with the students' biographical data, language proficiency and learning preferences, which will be collected one month prior to the experiment.

Although within-subject designs, and mixed designs for that matter, remove personal variability, they may be prone to influences originating from the materials used in the experiment. Therefore, meticulous preparation, preferably including pre-testing, and analysis of the materials is required. The aim of this paper is to show how this may be carried out.

3.2. The Ten Steps

A number of preparatory steps need to be taken to ensure the use of comparable materials in a within-subject AVTCRP study to safeguard the validity of future results. Based on our own experiences, we propose to divide the initial process of preparation into the ten distinct steps listed below:

(a) Careful preparation of materials
(b) Lecture content and feature analyses
(c) First pilot study
(d) Re-evaluation
(e) Optimization
(f) Second pilot study
(g) Production of comparable subtitles
(h) Subtitles analyses
(i) Third pilot study with subtitles
(j) Finalisation of materials

In the following paragraphs, the first six steps will be explained in detail, integrating relevant research. Each step will generate results which (if applicable) might be carried over and integrated into the next step. Given the limited scope of this article, we will focus exclusively on the preparation of the lectures (which can be considered source texts) and the comprehension tests (steps a–f). The complex production, analyses and testing of comparable interlingual and intralingual subtitles are beyond the scope of this paper and will be published in a future article.

3.2.1. Careful Preparation of Materials

Comparable materials are of the utmost importance for a within-subject design. In the S4AE project we examine the effect of no subtitles, interlingual (Dutch) subtitles and intralingual (English) subtitles. This implies we need three lectures that are comparable in content and language (complexity), length, style, etc. Content-wise, all three lectures focused on philosophy, which was realistic and viable, since optional courses in philosophy are part of the study program of the intended participants. Professor Frank Albers, philosophy lecturer at the University of Antwerp, wrote three comparable lectures on the views on inequality of three renowned philosophers, Thomas Piketty, Jean-Jacques Rousseau and Alexis de Tocqueville. The lecture texts were subsequently analysed and recorded (see Section 3.2.2).

In addition to the lecture texts, the measurement tools had to be selected and prepared. We used eye tracking and an (existing) psychometric self-report questionnaire to measure cognitive load (Leppink & van den Heuvel, 2015). This validated questionnaire consisted of eight general questions for which each participant had to rate complexity on a scale from 1 to 10, 1 representing low

complexity and 10 representing high complexity. The first four questions asked about content complexity and as such provided insight into the overall perceived intrinsic load. The last four concerned instructional complexity and thus provided data on perceived extraneous load. To measure retention, we used a (repeated) comprehension test. This tool had frequently been used successfully in earlier AVTCRP research (e.g., Lavaur & Bairstow, 2011; Montero Perez et al., 2014). We designed the comprehension tests as if they were exams for a philosophy course. All three tests consisted of twelve questions and had equal numbers of multiple-choice questions, input questions, memory questions and insight questions[2]. Finally, we used a biographical survey and would employ additional tests in the main experiment, e.g., proficiency tests aimed at assessing listening and reading competences in both English and Dutch and supplementary surveys, to accurately examine the participants' profiles, proficiency level and prior knowledge.

3.2.2. Lecture Content and Feature Analyses

The lecture texts were first compared in terms of readability to ensure their comparability[3]. To this end, we used the Flesch Reading Ease, the Flesch-Kincaid Grade Level and the New Dale-Chall. The first two calculate readability based on the average sentence length and the average number of syllables per word. The Flesch Reading Ease gives a score out of 100, for which above 90 is considered very easy and below 30 is considered very hard; the Flesch-Kincaid Grade Level indicates the American grade-school level necessary to be able to read the text. Sentence and word length are considered accurate indicators of readability (Smeuninx, 2018), but to include different measures, we also chose to add the New Dale-Chall formula, which calculates readability based on a list of familiar words and the average sentence length and gives a score ranging from 0 to 10 or above corresponding to a grade level (Table 1 reports the grade level). As shown in Table 1, the texts receive very similar scores and are estimated to be difficult texts aimed at twelfth grade (17–18yo) students.

Table 1.

Text Readability Scores

	Piketty (P1)	Rousseau (R2)	Tocqueville (T3)
Flesch Reading Ease	43	41	38
Flesch-Kincaid Grade Level	12	12	12
New Dale-Chall	11-12	11-12	11-12

[2] We have not released the comprehension tests yet, as they will still be used in various experiments. Please contact the author for a confidential copy if desired.

[3] Although the experimental order was completely counterbalanced, the order in which the lectures are discussed in this article will, for consistency, always be Piketty first (referred to as P1), Rousseau second (R2) and de Tocqueville last (T3).

We then analysed the texts using Perego et al.'s (2018) construct for film complexity. These researchers distinguish three types of complexity: (a) structural-informative complexity, i.e., number of cuts as a measure of newly introduced information, pace and total number of one and two-line subtitles, (b) linguistic complexity, i.e., total word count, standardised type-token ratio (TTR), words per minute (WPM), total sentence count and average sentence length, and (c) narrative complexity, i.e., number of film locations, number of characters and number of flashbacks. Structural-informative complexity is not relevant at this stage given the absence of subtitles and cuts. Table 2 shows the relevant indices for linguistic complexity, with the word count and standardised TTR being very similar. Sentence count and length vary, but this is deemed less important as these texts will be recorded as lectures (oral texts). WPM/WPS is discussed below (Table 3). Perego et al. (2018) mention chronology and amount of information as key aspects of narrative complexity. After analysing the texts, we concluded similar information was presented in a comparable order.

Table 2.

Linguistic Complexity

	Piketty (P1)	Rousseau (R2)	Tocqueville (T3)
Total word count	797	798	800
Standardised TTR	0.517	0.451	0.469
Total sentence count	48	56	64
Average sentence length in words	16.604	14.25	12.5

The lectures were subsequently recorded in a recording studio using an identical format. In each of the three lecture recordings, Professor Albers is shown against a black background. This talking head format is, of course, a more artificial setting than a normal classroom environment, i.e., lower external validity, but the research project aims to assess the impact of subtitles in a more controlled environment. Additionally, minimising the effects of the lecturer also reduces extraneous load and increases information transfer following the coherence effect (Mayer & Moreno, 2003). This may enable the students to read and process the subtitles better, which has been shown to correlate directly with performance (Kruger & Steyn, 2014).

Finally, the lecture recordings were analysed. Each lecture is approximately 7 minutes long. The professor does not use hand gestures nor does he cough, he has a constant intonation, rarely stutters and has a relatively constant facial expression across all three lectures. One notable difference from the lecture texts is that the professor tends to explicitly mention quotation marks or add various expressions for indirect speech to signalise quotes. This results in a slightly different total number of words in the lecture. Table 3 shows the length of each recording, the adjusted word count, the overall speech rate in words per second (WPS) and the mean speech rate across 14 intervals of 30 seconds in WPS. Based on these aspects, our team considered the lecture recordings comparable.

Table 3.

Lecture Recordings Indices

	Piketty (P1)	Rousseau (R2)	Tocqueville (T3)
Recording Length (mm:ss)	7:21	7:08	7:25
Adjusted Word Count	833	855	833
Overall Speech Rate (WPS)	1.9463	2.0357	1.9018
Mean Interval Speech Rate (WPS)	1.9500	2.0357	1.9191

3.2.3. First Pilot Study

To verify the conclusions drawn from step 2, a first pilot study without subtitles was set up and conducted in May 2018 with 75 2nd-year students of the BA in Applied Linguistics from the University of Antwerp. They all completed the biographical survey, self-report cognitive load questionnaires and the comprehension tests. Eye tracking, pre-testing and post-testing were excluded to focus on the materials themselves and to keep data analysis feasible. For the statistical analyses, we have one within-subject variable with three levels, i.e., the three lectures, and two independent between-group variables with two levels: the study of English (i.e., studying English in their BA or not) and prior knowledge of philosophy (i.e., having followed an optional philosophy course taught by the professor featured in the lectures or not)[4]. We consistently use mixed ANOVAs as these can compare the mean differences between the lectures and take into account the two between-group variables. However, it is important to note that these between-group variables only provide rough indications of the students' profiles based on the biographical survey since extensive pre-testing (which will be done in the main study) was foregone at this stage. Consequently, we mainly focus on the within-subject effects for all participants and will only briefly discuss interactions with these between-group variables.

A number of conclusions could be drawn from this experiment[5]:

Firstly, T3 appears to induce significantly lower total load (mean of all questions in the psychometric self-report) than P1 and R2 for all participants (Appendix, Table 5.3). The same can be observed for intrinsic load (mean of questions 1-4; Appendix, Table 6.3). In contrast, no significant main effects were found for extraneous load (mean of questions 5–8; Appendix, Table 7.2). As far as interaction effects are concerned, we observed a significant interaction effect between total load and between-group philosophy variable (Appendix, Table 5.2), and between extraneous load ratings and

[4] The exact participant distributions for each of these variables and the relevant mean ratings and scores can be found in the Appendix, Tables 5.1, 6.1, 7.1 and 8.1.

[5] Due to space constraints, extensive reporting and statistics were omitted in this section but can be found in the Appendix, Tables 5–8.

philosophy (Appendix, Table 7.2). In terms of between-group effects, those studying English show significantly lower total load ratings than those that do not (Appendix, Table 7.3).

For comprehension, we found a significant main effect of the lectures, but no significant interactions when the between-group variables are considered (Appendix, Table 8.2). It was revealed that participants scored significantly lower on R2 than on P1 or T3 (Appendix, Table 8.3). For the between-group effects, those studying philosophy were found to perform better than those that did not (Appendix, Table 8.4).

In this pilot study, we were mainly interested in the differences regardless of groups, which explains why the comprehension results are particularly problematic. These tests need revising since the lack of comparability might not reside in the lectures but in the comprehension questions themselves. In this light, the overall difference in total load, and consequently intrinsic load, between T3 and the other lectures may also be problematic, since it might indeed hint at a difference between the lectures. However, we believe that data noise could be an issue. By data noise we mean the data produced by participants who did not follow the instructions properly[6], e.g., a participant rating all psychometric questions with the same number just to be done with the experiment or always choosing the first multiple-choice answer in the comprehension tests. We did not verify whether the participants actually watched the videos or followed the instructions and were therefore unable to filter this possibly conflicting, inaccurate or meaningless data. Accordingly, we will first focus on the revision of the comprehension tests and implement some sort of participant surveillance.

3.2.4. Re-evaluation

Following the results from the first pilot study, all materials were re-evaluated in an attempt to pinpoint a possible cause for the differences. Our team of researchers unanimously agreed that, although T3 could be considered slightly easier content-wise due to it being less philosophical and more focused on political rather than monetary (in)equality, the main problem resided in the comprehension tests and the lack of data noise prevention. Consequently, the need for optimization of the comprehension tests arose.

3.2.5. Optimization

We recomposed the tests in view of our within-subject component. We no longer focused on creating tests similar to actual lecture exams, but instead aimed to strengthen comparability between questions for all lectures, including not only main ideas but also secondary details. Due to a lack of

[6] Not to be confused with noise in eye tracking data which refers to data being unusable due to signal loss, inaccuracy of the eye tracker, etc.

research on how to develop comparable within-subject comprehension tests, we devised our own approach. First, all originally used questions were considered, disregarding scores, to establish so-called matches (i.e., comparable questions across the three tests), using a large number of variables such as question type, answer type, question length, answer length, in-text location of the first mention of the answer, in-text repetition of the answer and "hearing guesses" (i.e., the probability of guessing correctly based on listening to the lecture). If no match could be found for a particular question, it was discarded. If a match could be found between two lectures only, we explored the possibility of creating a similar question for the remaining lecture.[7] Consequently, each test contained twelve questions comparable to the questions in the other two tests. Although this may have eliminated undesirable influences from varying degrees of difficulty in the comprehension tests, we expect a possible increase in order and/or carryover effects (Mellinger & Hanson, 2017) and will verify this in statistical analyses. Lastly, we logged mouse activity to check whether participants watched the entire video and monitored participants more closely to prevent inattentive behaviour.

3.2.6. Second Pilot Study Without Subtitles

To test the optimised comprehension tests, we conducted a second pilot study without subtitles in March 2019 with 50 2nd-year students of the BA in Applied Linguistics of the University of Antwerp (33 female; 17 male)[8]. The same within-subject (the lectures) and between-group (English and philosophy) variables from the first pilot study were used. The participants filled in the biographical survey first. Then they watched the three lectures, each time followed by filling in the psychometric questionnaire (Leppink & van den Heuvel, 2015) and the respective comprehension test. As in the first pilot study, mixed ANOVAs were used to analyse the data.

The mean total, intrinsic and extraneous load were relatively similar for all three lectures (Appendix, Tables 9.1, 10.1 and 11.1). Additionally, the average scores for the three types of cognitive load for each lecture individually were very similar to the scores from the first pilot study.

We first analysed the within-subject effects for total load (Appendix, Table 9.1). Maulchy's Test of Sphericity confirmed spherical data, $X^2(2) = 2.879$, $p = 0.237$; a mixed ANOVA only found a significant interaction effect with the English variable, $F(2, 86) = 5.234$, $p = 0.007$, but no significant main effects were found for all participants, $F(2, 86) = 2.808$, $p = 0.066$ (Appendix, Table 9.2). No between-group effects were revealed either (Appendix, Table 9.3).

For intrinsic load (Appendix, Table 10.1), Maulchy's Test of Sphericity revealed a violation of sphericity, $X^2(2) = 9.018$, $p = 0.011$. With a Greenhouse-Geisser correction for non-spherical data, a

[7] Due to space constraints, in-depth explanations of the question categorisation were not included in this paper. If necessary, contact the author for more information.

[8] Distributions in Appendix Tables may vary as some scores or ratings were excluded following the data noise filtering.

mixed ANOVA showed no statistically significant main within-subject effect of the lectures on intrinsic load for all participants, $F(1.676, 72.074) = 2.913$, $p = 0.070$, and no interaction effects (Appendix, Table 10.2). Furthermore, no significant between-group effects were found (Appendix, Table 10.3).

Lastly, we looked at extraneous load (Appendix, Table 11.1). After assuming sphericity, as Maulchy's test of sphericity revealed spherical data, $X^2(2) = 5.082$, $p = 0.079$, a mixed ANOVA indicated that the lectures did not significantly differ for all participants in extraneous load either, $F(2, 86) = 1.581$, $p = 0.212$ (Appendix, Table 11.2). Furthermore, it only showed a significant interaction between the extraneous load ratings and the English variable, $F(2, 86) = 6.567$, $p = 0.002$ (Appendix, Table 11.2). We consider this interaction of the English variable not to be problematic, since more extensive testing of the proficiency level will be done for the main experiment and the extraneous load ratings regardless of groups do not differ significantly. In terms of between-group effects, those studying English perceived a significantly lower extraneous load across the videos, $F(1, 43) = 9.414$, $p = 0.004$, $r = 0.47$, than the others. A significantly lower extraneous load was also found for the philosophy students when compared to those who did not follow any philosophy course, $F(1, 43) = 5.535$, $p = 0.023$, $r = 0.36$ (Appendix, Table 11.3). It could be expected that the English students had fewer problems with a course taught in English (extraneous load) given their proficiency level. Experience with philosophy on the other hand was expected to have an effect on content comprehension (intrinsic load) instead of extraneous load, which was not found.

In contrast to the findings of the first pilot study, we no longer found a difference between the cognitive loads of the three lectures for all participants. However, we found two significant interaction effects with the between-group English variable.

We draw two conclusions from these findings. Firstly, the importance of participant surveillance and double-checking mechanisms to verify the viewing and proper answering of the questionnaires and tests cannot be underestimated. The significant main effect of the lectures on total and intrinsic load in the first pilot study disappeared in the second pilot study. The contrast between these findings, with data noise filtering being the only difference, is striking. Secondly, in similar within-subject studies it is key to accurately assess participant profiles, prior knowledge and language proficiency.

With regard to the comprehension scores, R2 again received the lowest mean score and T3 the highest, with P1 scoring in between (Appendix, Table 12.1). However, a mixed ANOVA revealed no significant main within-subject effects for all participants and no interaction effects (Appendix, Table 12.2). The optimisation of the comprehension tests for this particular within-subject experiment clearly helped. Both the cognitive load ratings and comprehension test scores indicate that the lectures and the comprehension tests are comparable.

Despite these already promising results, we decided to improve the comprehension test even further to flatten out minor insignificant differences that may still be present between the tests. Based on several guidelines (Demeuse & Henry, 2004; Professional Testing, 2020), advice from statisticians from the University of Antwerp on test item analyses and parts of the Item Response Theory (Baker,

2001), we decided to disregard a number of questions in the test. Three variables were used to decide which questions to disregard: difficulty, discrimination and reliability.

The difficulty score of a question is based on the percentage of examinees having answered that question correctly. Since the threshold for what is considered to be a difficult or an easy question is arbitrary, we adhere to the guidelines of the University of Antwerp: if less than 10% of the participants answer correctly, the question is considered to be difficult, whereas a question is easy when more than 90% answer it correctly. Questions that are too difficult or too easy would no longer be considered in future testing. The discrimination score reveals whether a question is in line with what is assessed, assuming that an examinee with high overall testing scores has a higher chance of answering a question correctly. If a question tends to be answered correctly more often by examinees who obtain lower overall scores, while the better examinees tend to answer that question incorrectly, that question can be considered to not be discriminating and not in line with what is assessed. In our university guidelines, the discrimination score for each question is calculated by deducting the number of correct answers in the worst scoring 25% of the participants from the number of correct answers in the best scoring 25% of the students and dividing that number by the largest of those two numbers. It is advised that questions with discrimination scores lower than 0.20 are disregarded in future testing. Our university guidelines determine reliability/consistency with the Pearson point-biserial correlation coefficient between the question and the total scores and should ideally be equal to or higher than 0.15. Similar to the discrimination score, this variable reveals whether the question is in line with what the entire test wants to assess.

When a question was flagged for two of the three variables, we decided to disregard the question in further analyses. We chose to exempt questions instead of discarding them to maintain an equal number of questions and safeguard the similarities between the comprehension tests for the three lectures. This eventually led to three twelve-question tests, but for P1 the scores of only 10 distinct questions were considered, for R2 10 questions and for T3 11 questions. When we compared the newly weighted average scores of the three tests (Appendix, Table 13.1), we see highly similar scores. After verifying sphericity with Maulchy's test of sphericity, $X^2(2) = 2.503$, $p = 0.286$, a mixed ANOVA again revealed no significant main within-subject effect for all participants, $F(2, 88) = 0.469$, $p = 0.627$, and no interaction effects (Appendix, Table 13.2).

Although this additional enhancement was not required, it clearly strengthened the similarity of the tests in terms of test scores and can be used as another example of adjusting the materials to benefit comparability in within-subject studies.

Since Mellinger and Hanson (2018, p. 16) warned that there might still be order effects despite having counterbalanced orders, we checked whether psychometric ratings, comprehension scores and recoded comprehension scores for each lecture differed when it was watched first compared to when it was watched second or third. No real pattern in cognitive load or comprehension could be detected (see Table 4).

Table 4.

Second Pilot Study – Mean Cognitive Load Ratings & Comprehension Scores Based on Order

	Place in series	Piketty (P1)	Rousseau (R2)	Tocqueville (T3)
Intrinsic Load	1	6.2206	5.5694	4.9667
	2	5.8214	6.0294	5.1471
	3	5.3750	5.1667	5.8438
Extraneous Load	1	3.0000	3.4167	2.6000
	2	3.0357	3.1029	3.1324
	3	3.5000	2.9167	3.0156
Total Load	1	4.6103	4.4931	3.7833
	2	4.4286	4.5662	4.1397
	3	4.4375	4.0417	4.4297
Comprehension	1	51.18%	47.22%	46.06%
	2	44.29%	51.76%	55.08%
	3	53.53%	50.00%	48.30%

4. Conclusions

The use of within-subject designs is rather scarce in the body of research into AVT cognition, reception and perception. The aim of this article is not to plea for within-subject studies to take over the world of AVTCRP research. As Bernardini did in 2001 for TAP-based research, we advocate for more frequent use of within-subject designs in AVT research, conducted alongside between-subject studies. Within-subject designs could give additional insight into the idiosyncratic nature of perception, cognition and comprehension of AVT, and could increase statistical power in studies with limited sample sizes.

However, to safeguard validity, careful preparation and pre-testing of research materials and experimental set-up (preferably using at least two pilot studies) is key. A within-subject design might minimize characteristic influences due to the same participants being tested, but it also has a higher risk of undesirable influences from the materials or experimental setup. In this paper, we proposed a ten-step preparation of a within-subject AVTCRP study, which may guide or inspire future research. Based on the experience gained in the S4AE project, we can conclude it is rather challenging to develop materials and tools that are comparable in content and language (complexity), style, length, etc. We have also demonstrated the necessity to be cautious of initial subjective or intuitive assessment of comparability and to pre-test materials and measurement tools using objective measures. Additionally, we have shown that, instead of creating new materials or refurbishing measurement tools, there are other options to allow for valid and reliable results, for example, by recoding comprehension test scores based on an approach from educational research. Methodological input from the aforementioned field of education, other domains of Translation

Studies (e.g., translation process research) or even other fields (e.g., Psychology) may be useful to guide this pre-testing phase.

We acknowledge that the ten-step proposal needs adaptation dependent on specific research goals, as well as further refinement. We acknowledge limitations in our approach, such as potential bias in the initial preparatory steps and the relatively small participant sub-groups, particularly in the second pilot study. However, we hope that this proposal will spark a disciplinary debate on the use of within-subject (or mixed) design in AVT research and the ways in which methodological preparations can be approached.

Acknowledgements

We would like to thank Professor Frank Albers (University of Antwerp) for assisting us in the production of the lectures. We would also like to give our thanks to Professor Emerita Aline Remael for her help setting up the S4AE project. Lastly, we would like to express our gratitude to the statisticians of the University of Antwerp for their guidance regarding the recoding of the comprehension tests.

References

Bairstow, D. (2012). *Rôles des sous-titres dans la compréhension et la mémorisation de films* [The role of subtitles in the comprehension and retention of film] [Unpublished doctoral dissertation]. Paul Valéry University of Montpellier, France.

Bairstow, D., & Lavaur, J.-M. (2017). Sous-titrage, compréhension de films et acquisition de vocabulaire [Subtitles, comprehension of film and vocabulary acquisition]. *Psychologie Française, 62*(3), 249–261. https://doi.org/10.1016/j.psfr.2015.05.001

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

Balling, L. W., & Hvelplund, K. T. (2015). Design and statistics in quantitative translation (process) research. *Translation Spaces, 4*(1), 170–187. https://doi.org/10.1075/ts.4.1.08bal

Bernardini, S. (2001). Think-aloud protocols in translation research: Achievements, limits, future prospects. *Target, 13*(2), 241–263.

Bianchi, F., & Ciabattoni, T. (2008). Captions and subtitles in EFL learning: An investigative study in a comprehensive computer environment. In A. Baldry, M. Pavesi, C. T. Torsello, & C. Taylor (Eds.), *From didactas to ecolingua: An ongoing research project on translation and corpus linguistics* (pp. 69–90). Edizonioni Università di Trieste.

Birulés-Muntané, J., & Soto-Faraco, S. (2016). Watching subtitled films can help learning foreign languages. *PLoS One, 11*(6). https://doi.org/10.1371/journal.pone.0158409

Bisson, M.-J., Van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics, 35*(2), 399–418. https://doi.org/10.1017/s0142716412000434

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Colm, C. (2008). Using pupillometric, fixation-based and subjective measures to measure the processing effort experienced when viewing subtitled TV anime with pop-up gloss. *Copenhagen Studies in Language, 36,* 125–144.

Colm, C. (2009). *Relevant abuse? Investigating the effects of an abusive subtitling procedure on the perception of TV anime using eye tracker and questionnaire* [Unpublished doctoral dissertation]. Dublin City University, Ireland.

Danan, M. (1992). Reversed subtitling and dual coding theory: New directions for foreign language instruction. *Language Learning, 42*, 497–527. https://doi.org/10.1111/j.1467-1770.1992.tb01042.x

Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta: Journal des Traducteurs, 49*(1), 67–77. https://doi.org/10.7202/009021ar

de Linde, Z., & Kay, N. (1999). *The semiotics of subtitling*. St Jerome.

Demeuse, M., & Henry, G. (2004). L'analyse classique d'items [The classic analysis of items]. In M. Demeuse (Ed.), *Introduction aux théories et aux méthodes de la mesure en sciences psychologiques et en sciences de l'éducation* [Introduction to theories and methods of measurement in psychological and educational sciences] (pp. 173–186). Institut d'Administration scolaire de l'Université de Mons.

Díaz-Cintas, J. (2020). Audiovisual translation. In E. Angelone, M. Ehrensberger-Dow, & G. Massey (Eds.), *The Bloomsbury companion to language industry studies* (pp. 209–230). Bloomsbury Publishing.

Díaz-Cintas, J., & Cruz, M. F. (2008). Using subtitled video materials for foreign language instruction. In J. Díaz-Cintas (Ed.), *The didactics of audiovisual translation* (pp. 201–214). John Benjamins.

DíazCintas, J., & Szarkowska, A. (2020). Introduction: Experimental research in audiovisual translation – Cognition, reception, production. *The Journal of Specialised Translation, 33*, 3–16.

Doherty, S., & Kruger, J.-L. (2018). The development of eye tracking in empirical research on subtitling and captioning. In T. Dwyer, C. Perkins, S. Redmond, & J. Sita (Eds.), *Seeing into screens: Eye tracking and the moving image* (pp. 46–64). Bloomsbury.

d'Ydewalle, G., & De Bruycker, W. (2007). Eye movements of children and adults while reading television subtitles. *European Psychologist, 12*(3), 196–205. https://doi.org/10.1027/1016-9040.12.3.196

Frey, L. R., Botan, C. H., & Kreps, G. L. (1991). *Investigating communication: An introduction to research methods*. Pearson.

Gerber-Morón, O., & Szarkowska, A. (2018). Line breaks in subtitling: An eye tracking study on viewer preferences. *Journal of Eye Movement Research, 11*(3). https://doi.org/10.16910/jemr.11.3.2

Gerber-Morón, O., Szarkowska, A., & Woll, B. (2018). The impact of text segmentation on subtitle reading. *Journal of Eye Movement Research, 11*(4). https://doi.org/10.16910/jemr.11.4.2

Hefer, E. (2013a). Reading first and second language subtitles: Sesotho viewers reading in Sesotho and English. *Southern African Linguistics and Applied Language Studies, 31*(3), 359–373. https://doi.org/10.2989/16073614.2013.837610

Hefer, E. (2013b). Reading second language subtitles: A case study of Afrikaans viewers reading in Afrikaans and English. *Perspectives: Studies in Translation Theory and Practice, 21*(1), 22–41. https://doi.org/10.1080/0907676x.2012.722652

Jensema, C. J., El Sharkawy, S., Danturthi, R. S., Burch, R., & Hsu, D. (2000). Eye movement patterns of captioned television viewers. *American Annals of the Deaf, 145*(3), 275–285. https://doi.org/10.1353/aad.2012.0093

Kruger, J. (2013). Subtitles in the classroom: Balancing the benefits of dual coding with the cost of increased cognitive load. *Journal for Language Teaching, 47*(1), 55–71. https://doi.org/10.4314/jlt.v47i1.2

Kruger, J.-L., & Doherty, S. (2016). Measuring cognitive load in the presence of educational video: Towards a multimodal methodology. *Australasian Journal of Educational Technology, 32*(6), 19–31. https://doi.org/10.14742/ajet.3084

Kruger, J.-L., & Steyn, F. (2014). Subtitles and eye tracking: Reading and performance. *Reading Research Quarterly, 49*(1), 105–120. https://doi.org/10.1002/rrq.59

Kruger, J.-L., Chan, W. S., & Doherty, S. (2019). Comparing the impact of automatically generated and corrected subtitles on cognitive load and learning in a first- and second-language educational context. *Linguistica Antverpiensia, New Series: Themes in Translation Studies,* (18), 237–272.

Kruger, J.-L., Doherty, S., Fox, W., & de Lissa, P. (2018). Multimodal measurement of cognitive load during subtitle processing: Same-language subtitles for foreign-language viewers. In I. Lacruz & R. Jääskeläinen (Eds.), *Innovation and expansion in translation process research* (pp. 267–294). John Benjamins.

Kruger, J.-L., Hefer, E., & Matthew, G. (2013, August 29–31). *Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts* [Conference presentation]. The 2013 Conference on Eye Tracking South Africa, Cape Town, South Africa.

Kruger, J.-L., Hefer, E., & Matthew, G. (2014). Attention distribution and cognitive load in a subtitled academic lecture: L1 vs. L2. *Journal of Eye Movement Research, 7*(5), 1–15. https://doi.org/10.16910/jemr.7.5.4

Kruger, J.-L., Soto-Sanfiel, M. T., Doherty, S., & Ibrahim, R. (2016). Towards a cognitive audiovisual translatology. In R. M. Martín (Ed.), *Reembedding translation process research* (pp. 171–193). John Benjamins.

Kruger, J.-L., Szarkowska, A., & Krejtz, I. (2015). Subtitles on the moving image: An overview of eye tracking studies. *Refractory: A Journal of Entertainment Media, 25*, 1–14.

Lavaur, J.-M., & Bairstow, D. (2011). Languages on the screen: Is film comprehension related to the viewers' fluency level and to the language in the subtitles? *International Journal of Psychology, 46*(6), 455–462. https://doi.org/10.1080/00207594.2011.565343

Leppink, J., & van den Heuvel, A. (2015). The evolution of cognitive load theory and its application to medical education. *Perspectives on Medical Education, 4*(3), 119–127. https://doi.org/10.1007/s40037-015-0192-x

Liao, S., Kruger, J.-L., & Doherty, S. (2020). The impact of monolingual and bilingual subtitles on visual attention, cognitive load and comprehension. *The Journal of Specialised Translation, 33*, 70–98.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Mellinger, C. D., & Hanson, T. A. (2017). *Quantitative research methods in translation and interpreting studies*. Routledge.

Mellinger, C. D., & Hanson, T. A. (2018). Order effects in the translation process. *Translation, Cognition & Behavior, 1*(1), 1–20. https://doi.org/10.1075/tcb.00001.mel

Montero Perez, M. (2019). Pre-learning vocabulary before viewing captioned video: An eye tracking study. *The Language Learning Journal, 47*(4), 460–478. https://doi.org/10.1080/09571736.2019.1638623

Montero Perez, M. (2020). Incidental vocabulary learning through viewing video: The role of vocabulary knowledge and working memory. *Studies in Second Language Acquisition*, 1–25. https://doi.org/10.1017/s0272263119000706

Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology, 18*(1), 118–141.

Moreno, A. O. (2017). *Attention and dual coding theory: An interaction model using subtitles as a paradigm* [Unpublished doctoral dissertation]. Autonomous University of Barcelona, Spain.

Moreno, M. R., Otero-Sabogal, R., & Soto, C. (2011). Using web-based training to improve skills among bilingual dual-role staff interpreters. *International Journal of Interpreter Eduation, 3*, 22–48.

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*(1), 156–163. https://doi.org/10.1037/0022-0663.94.1.156

Orero, P., Doherty, S., Kruger, J.-L., Matamala, A., Pedersen, J., Perego, E., Romero-Fresco, P., Rovira-Esteva, S., Soler-Vilageliu, O., & Szarkowska, A. (2018). Conducting experimental research in audiovisual translation (AVT): A position paper. *The Journal of Specialised Translation, 30*, 105–126.

Orrego-Carmona, D. (2015). *The reception of (non)professional subtitling* [Unpublished doctoral dissertation]. University of Rovira I Virgili, Tarragona, Spain.

Perego, E., Del Missier, F., & Stragà, M. (2018). Dubbing vs. subtitling: Complexity matters. *Target, 30*(1), 137–157.

Perego, E., Del Missier, F., Porta, M., & Mosconi, M. (2010). The cognitive effectiveness of subtitle processing. *Media Psychology, 13*(3), 243–272. https://doi.org/10.1080/15213269.2010.502873

Perego, E., Laskowska, M., Matamala, A., Remael, A., Robert, I. S., Szarkowska, A., Vilaró, A., & Bottiroli, S. (2016a). Is subtitling equally effective everywhere? A first cross-national study on the reception of interlingually subtitled messages. *Across Languages and Cultures, 17*(2), 205–229. https://doi.org/10.1556/084.2016.17.2.4

Perego, E., Orrego-Carmona, D., & Bottiroli, S. (2016b). An empirical take on the dubbing vs. subtitling debate: An eye movement study. *Lingue e Linguaggi, 19*, 255–274. https://doi.org/10.1285/i22390359v19p255

Professional Testing. (2020). Step 9. Conduct the item analysis. http://www.proftesting.com/test_topics/pdfs/steps_9.pdf

Saldanha, G., & O'Brien, S. (2013). *Research methodologies in translation studies*. St Jerome.

Smeuninx, N. (2018). *Dear stakeholder: Exploring the language of sustainability reporting: A closer look at readability, sentiment and perception* [Unpublished doctoral dissertation]. Ghent University, Ghent, Belgium. https://biblio.ugent.be/publication/8576556

Szarkowska, A., & Bogucka, L. (2019). Six-second rule revisited: An eye tracking study on the impact of speech rate and language proficiency on subtitle reading. *Translation, Cognition & Behavior, 2*(1), 101–124. https://doi.org/10.1075/tcb.00022.sza

Szarkowska, A., & Gerber-Morón, O. (2018a). Viewers can keep up with fast subtitles: Evidence from eye movements. *PLoS One, 13*(6). https://doi.org/ 10.1371/journal.pone.0199331

Szarkowska, A., & Gerber-Morón, O. (2018b). Two or three lines: A mixed-methods study on subtitle processing and preferences. *Perspectives: Studies in Translation Theory and Practice, 27*(1). https://doi.org/10.1080/0907676X.2018.1520267

Szarkowska, A., Krejtz, I., Klyswejko, Z., & Wieczorek, A. (2011). Verbatim, standard, or edited? Reading patterns of different captioning styles among deaf, hard of hearing and hearing viewers. *American Annals of the Deaf, 156*(4), 363–378. https://doi.org/10.1353/aad.2011.0039

Tsaousi, A. (2016). *Sound effect labelling in subtitling for the deaf and hard-of-hearing: An experimental study on audience reception* [Unpublished doctoral dissertation]. Autonomous University of Barcelona, Spain.

Vulchanova, M., Aurstad, L. M. G., Kvitnes, I. E. N., & Eshuis, H. (2015). As naturalistic as it gets: Subtitles in the English classroom in Norway. *Frontiers in Psychology, 5*(1510). https://doi.org/10.3389/fpsyg.2014.01510

Wächter, B., & Maiworm, F. (2014). *English-taught programmes in European higher education: The state of play in 2014*. Lemmens.

Appendix

Table 5.

First Pilot – Total Load

5.1 - First Pilot Study - Descriptive Statistics					
Lecture	**English**	**Philosophy**	**Mean**	**Standard Deviation**	**N**
Piketty (P1)	No	No	5	1.57784	13
		Yes	5.3125	1.55373	10
		Total	5.1359	1.53968	23
	Yes	No	4.1477	1.20307	22
		Yes	4.3625	1.17689	30
		Total	4.2716	1.18112	52
	Total	No	4.4643	1.39543	35
		Yes	4.6	1.32687	40
		Total	4.5367	1.35173	75
Rousseau (R2)	No	No	5.5481	1.54071	13
		Yes	4.4375	0.98116	10
		Total	5.0652	1.41616	23
	Yes	No	4.5455	0.99892	22
		Yes	4.1875	1.54834	30
		Total	4.3389	1.34386	52
	Total	No	4.9179	1.3022	35
		Yes	4.25	1.42015	40
		Total	4.5617	1.39805	75
Tocqueville (T3)	No	No	4.3269	1.18973	13
		Yes	4.325	1.41446	10
		Total	4.3261	1.26117	23
	Yes	No	3.9489	0.96413	22
		Yes	4.0583	1.36881	30
		Total	4.012	1.20463	52
	Total	No	4.0893	1.05265	35
		Yes	4.125	1.36696	40
		Total	4.1083	1.22239	75

5.2 - First Pilot Study - Test of Within-Subject Effects							
Source		**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Total	Sphericity Assumed	11.645	2	5.822	5.87	0.004*	0.076
*Total * English*	Sphericity Assumed	2.622	2	1.311	1.322	0.27	0.018
*Total * Philosophy*	Sphericity Assumed	8.66	2	4.33	4.366	0.014*	0.058
*Total * English * Philosophy*	Sphericity Assumed	1.536	2	0.768	0.774	0.463	0.011
Error (Intrinsic)	Sphericity Assumed	140.845	142	0.992			

5.3 - First Pilot Study - Post-Hoc Pairwise Comparisons (Bonferroni)						
Total		**Mean Difference**	**Standard Error**	**Significance**	**95% Confidence Interval for Difference**	
					Lower Bound	Upper Bound
Piketty (P1)	Rousseau (R2)	0.026	0.176	1	-0.407	0.459
	Tocqueville (T3)	0.541	0.156	0.003*	0.158	0.924
Rousseau (R2)	Piketty (P1)	-0.026	0.176	1	-0.459	0.407
	Tocqueville (T3)	0.515	0.199	0.035*	0.027	1.003
Tocqueville (T3)	Piketty (P1)	-0.541	0.156	0.003*	-0.924	-0.158
	Rousseau (R2)	-0.515	0.199	0.035*	-1.003	-0.027

11.4 - First Pilot Study - Test of Between-Subject Effects						
Source	**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intercept	1276.476	1	1276.476	1248.379	0.000*	0.946
English	5.947	1	5.947	5.816	0.018*	0.076
Philosophy	0.302	1	0.302	0.295	0.588	0.004
*English * Philosophy*	0.255	1	0.255	0.25	0.619	0.004
Error	72.598	71	1.023			

Table 6.

First Pilot – Intrinsic Load

6.1 - First Pilot Study - Descriptive Statistics					
Lecture	***English***	***Philosophy***	***Mean***	***Standard Deviation***	***N***
Piketty (P1)	*No*	*No*	*5.8077*	*1.54837*	*13*
		Yes	*6.725*	*1.41151*	*10*
		Total	*6.2065*	*1.52936*	*23*
	Yes	*No*	*5.4091*	*1.49114*	*22*
		Yes	*5.6*	*1.20631*	*30*
		Total	*5.5192*	*1.32366*	*52*
	Total	*No*	*5.5571*	*1.50255*	*35*
		Yes	*5.8813*	*1.33612*	*40*
		Total	*5.73*	*1.41586*	*75*
Rousseau (R2)	*No*	*No*	*6.2885*	*1.45361*	*13*
		Yes	*5.775*	*1.62211*	*10*
		Total	*6.0652*	*1.51548*	*23*
	Yes	*No*	*5.75*	*1.20515*	*22*
		Yes	*5.375*	*1.83212*	*30*
		Total	*5.5337*	*1.59428*	*52*
	Total	*No*	*5.95*	*1.30863*	*35*
		Yes	*5.475*	*1.7703*	*40*
		Total	*5.6967*	*1.57969*	*75*
Tocqueville (T3)	*No*	*No*	*5.1346*	*1.51594*	*13*
		Yes	*5.375*	*1.7129*	*10*
		Total	*5.2391*	*1.57119*	*23*
	Yes	*No*	*5.1364*	*1.24098*	*22*
		Yes	*4.9917*	*1.59131*	*30*
		Total	*5.0529*	*1.44196*	*52*
	Total	*No*	*5.1357*	*1.32751*	*35*
		Yes	*5.0875*	*1.60882*	*40*
		Total	*5.11*	*1.47458*	*75*

6.2 - First Pilot Study - Test of Within-Subject Effects							
Source		**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intrinsic	Sphericity Assumed	19.639	2	9.819	7.479	0.001*	0.095
*Intrinsic * English*	Sphericity Assumed	2.551	2	1.275	0.971	0.381	0.013
*Intrinsic * Philosophy*	Sphericity Assumed	7.796	2	3.898	2.969	0.055	0.04
*Intrinsic * English * Philosophy*	Sphericity Assumed	1.484	2	0.742	0.565	0.57	0.008
Error (Intrinsic)	Sphericity Assumed	186.439	142	1.313			

6.3 - First Pilot Study - Post-Hoc Pairwise Comparisons (Bonferroni)						
Intrinsic		***Mean Difference***	***Standard Error***	***Significance***	***95% Confidence Interval for Difference***	
					Lower Bound	*Upper Bound*
Piketty (P1)	*Rousseau (R2)*	*0.088*	*0.189*	*1*	*-0.374*	*0.551*
	Tocqueville (T3)	*0.726*	*0.191*	*0.001**	*0.257*	*1.195*
Rousseau (R2)	*Piketty (P1)*	*-0.088*	*0.189*	*1*	*-0.551*	*0.374*
	Tocqueville (T3)	*0.638*	*0.232*	*0.023**	*0.069*	*1.206*
Tocqueville (T3)	*Piketty (P1)*	*-0.726*	*0.191*	*0.001**	*-1.195*	*-0.257*
	Rousseau (R2)	*-0.638*	*0.232*	*0.023**	*-1.206*	*-0.069*

6.4 - First Pilot Study - Test of Between-Subject Effects						
Source	**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intercept	*1972.032*	*1*	*1972.032*	*1469.735*	*0.000**	*0.954*
English	*3.514*	*1*	*3.514*	*2.619*	*0.11*	*0.036*
Philosophy	*0.043*	*1*	*0.043*	*0.032*	*0.858*	*0*
*English * Philosophy*	*0.411*	*1*	*0.411*	*0.307*	*0.582*	*0.004*
Error	*95.265*	*71*	*1.342*			

Table 7.

First Pilot – Extraneous Load

7.1 - First Pilot Study - Descriptive Statistics					
Lecture	**English**	**Philosophy**	**Mean**	**Standard Deviation**	**N**
Piketty (P1)	No	No	4.1923	2.01854	13
		Yes	3.9	2.322	10
		Total	4.0652	2.10953	23
	Yes	No	2.8864	1.56739	22
		Yes	3.125	1.69526	30
		Total	3.024	1.63094	52
	Total	No	3.3714	1.83449	35
		Yes	3.3188	1.86996	40
		Total	3.3433	1.84115	75
Rousseau (R2)	No	No	4.8077	2.06447	13
		Yes	3.1	1.51932	10
		Total	4.0652	2.00456	23
	Yes	No	3.3409	1.52469	22
		Yes	3	1.90621	30
		Total	3.1442	1.74709	52
	Total	No	3.8857	1.85934	35
		Yes	3.025	1.79904	40
		Total	3.4267	1.86575	75
Tocqueville (T3)	No	No	3.5192	1.58266	13
		Yes	3.275	1.52046	10
		Total	3.413	1.52556	23
	Yes	No	2.7614	1.46685	22
		Yes	3.125	1.92729	30
		Total	2.9712	1.74098	52
	Total	No	3.0429	1.53331	35
		Yes	3.1625	1.81655	40
		Total	3.1067	1.68016	75

7.2 - First Pilot Study - Test of Within-Subject Effects							
Source		**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Extraneous	Sphericity Assumed	5.872	2	2.936	2.369	0.097	0.032
*Extraneous * English*	Sphericity Assumed	2.704	2	1.352	1.091	0.339	0.015
*Extraneous * Philosophy*	Sphericity Assumed	11.354	2	5.677	4.58	0.012*	0.061
*Extraneous * English * Philosophy*	Sphericity Assumed	1.669	2	0.835	0.673	0.512	0.009
Error (Intrinsic)	Sphericity Assumed	176.009	142	1.24			

7.3 - First Pilot Study - Test of Between-Subject Effects						
Source	**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intercept	731.597	1	731.597	320.49	0.000*	0.819
English	9.018	1	9.018	3.95	0.051	0.053
Philosophy	1.708	1	1.708	0.748	0.39	0.01
*English * Philosophy*	2.728	1	2.728	1.195	0.278	0.017
Error	162.075	71	2.283			

Table 8.

First Pilot – Comprehension

8.1 - First Pilot Study - Descriptive Statistics (Absolute Score)					
Lecture	**English**	**Philosophy**	**Mean**	**Standard Deviation**	**N**
Piketty (P1)	No	No	6.08	1.706	13
		Yes	7.1	2.079	10
		Total	6.52	1.904	23
	Yes	No	6.18	1.967	22
		Yes	6.17	2.019	30
		Total	6.17	1.978	52
	Total	No	6.14	1.849	35
		Yes	6.4	2.048	40
		Total	6.28	1.949	75
Rousseau (R2)	No	No	4.08	0.862	13
		Yes	5.1	2.025	10
		Total	4.52	1.534	23
	Yes	No	5.23	1.744	22
		Yes	5.93	2.164	30
		Total	5.63	2.01	52
	Total	No	4.8	1.568	35
		Yes	5.73	2.136	40
		Total	5.29	1.937	75
Tocqueville (T3)	No	No	6	2.198	13
		Yes	6.3	2.83	10
		Total	6.13	2.437	23
	Yes	No	6.86	1.833	22
		Yes	7.6	1.976	30
		Total	7.29	1.934	52
	Total	No	6.54	1.99	35
		Yes	7.28	2.253	40
		Total	6.93	2.152	75

8.2 - First Pilot Study - Test of Within-Subject Effects							
Source		**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Comprehension	Greenhouse-Geisser	90.913	1.707	53.269	12.419	0.000*	0.149
*Comprehension * English*	Greenhouse-Geisser	22.021	1.707	12.903	3.008	0.061	0.041
*Comprehension * Philosophy*	Greenhouse-Geisser	1.305	1.707	0.764	0.178	0.803	0.003
*Comprehension * English * Philosophy*	Greenhouse-Geisser	4.252	1.707	2.492	0.581	0.535	0.008
Error (Intrinsic)	Greenhouse-Geisser	519.755	121.174	4.289			

8.3 - First Pilot Study - Post-Hoc Pairwise Comparisons (Bonferroni)						
Comprehension		**Mean Difference**	**Standard Error**	**Significance**	**95% Confidence Interval for Difference**	
					Lower Bound	Upper Bound
Piketty (P1)	Rousseau (R2)	1.297	0.353	0.001*	0.431	2.163
	Tocqueville (T3)	-0.31	0.393	1	-1.274	0.654
Rousseau (R2)	Piketty (P1)	-1.297	0.353	0.001*	-2.163	-0.431
	Tocqueville (T3)	-1.607	0.268	0.000*	-2.263	-0.95
Tocqueville (T3)	Piketty (P1)	0.31	0.393	1	-0.654	1.274
	Rousseau (R2)	1.607	0.268	0.000*	0.95	2.263

8.4 - First Pilot Study - Test of Between-Subject Effects						
Source	**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intercept	2291.919	1	2291.919	1559.916	0.000*	0.956
English	4.786	1	4.786	3.258	0.075	0.044
Philosophy	6.187	1	6.187	4.211	0.044*	0.056
*English * Philosophy*	0.367	1	0.367	0.25	0.619	0.004
Error	104.317	71	1.469			

Table 9.

Second Pilot – Total Load

9.1 - Second Pilot Study - Descriptive Statistics					
Lecture	**English**	**Philosophy**	**Mean**	**Standard Deviation**	**N**
Piketty (P1)	No	No	5.4583	0.98027	9
		Yes	5.0556	2.07017	9
		Total	5.2569	1.5849	18
	Yes	No	4.2989	1.82343	23
		Yes	2.9792	1.11079	6
		Total	4.0259	1.76883	29
	Total	No	4.625	1.69944	32
		Yes	4.225	1.99955	15
		Total	4.4973	1.78849	47
Rousseau (R2)	No	No	4.7639	1.07972	9
		Yes	4.3889	2.46969	9
		Total	4.5764	1.85907	18
	Yes	No	4.538	1.40991	23
		Yes	3.2917	0.32275	6
		Total	4.2802	1.35812	29
	Total	No	4.6016	1.31233	32
		Yes	3.95	1.95759	15
		Total	4.3936	1.55601	47
Tocqueville (T3)	No	No	4.2917	1.29452	9
		Yes	4.3333	2.03869	9
		Total	4.3125	1.65679	18
	Yes	No	4.0978	1.70684	23
		Yes	3.375	0.67082	6
		Total	3.9483	1.56785	29
	Total	No	4.1523	1.5836	32
		Yes	3.95	1.66489	15
		Total	4.0878	1.5946	47

9.2 - Second Pilot Study - Test of Within-Subject Effects							
Source		**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Total	Sphericity Assumed	3.321	2	1.661	2.808	0.066	0.061
*Total * English*	Sphericity Assumed	6.19	2	3.095	5.234	0.007*	0.109
*Total * Philosophy*	Sphericity Assumed	1.525	2	0.763	1.29	0.281	0.029
*Total * English * Philosophy*	Sphericity Assumed	0.028	2	0.014	0.024	0.976	0.001
Error (Intrinsic)	Sphericity Assumed	50.856	86	0.591			

9.3 - Second Pilot Study - Test of Between-Subject Effects						
Source	**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intercept	665.07	1	665.07	300.181	0.000*	0.875
English	8.382	1	8.382	3.783	0.058	0.081
Philosophy	4.163	1	4.163	1.879	0.178	0.042
*English * Philosophy*	1.675	1	1.675	0.756	0.389	0.017
Error	95.269	43	2.216			

Table 10.

Second Pilot – Intrinsic Load

10.1 - Second Pilot Study - Descriptive Statistics					
Lecture	**English**	**Philosophy**	**Mean**	**Standard Deviation**	**N**
Piketty (P1)	No	No	6.1111	1.36994	9
		Yes	6.1667	1.82431	9
		Total	6.1389	1.5653	18
	Yes	No	5.6957	2.03235	23
		Yes	5.2917	2.19896	6
		Total	5.6121	2.03385	29
	Total	No	5.8125	1.85785	32
		Yes	5.8167	1.95591	15
		Total	5.8138	1.86826	47
Rousseau (R2)	No	No	5.2778	1.38318	9
		Yes	5.5278	2.35001	9
		Total	5.4028	1.87503	18
	Yes	No	5.9457	1.38981	23
		Yes	5.3333	1.21106	6
		Total	5.819	1.35768	29
	Total	No	5.7578	1.39914	32
		Yes	5.45	1.92075	15
		Total	5.6596	1.56943	47
Tocqueville (T3)	No	No	4.7222	1.38318	9
		Yes	5.6667	2.22907	9
		Total	5.1944	1.86405	18
	Yes	No	5.3152	1.84978	23
		Yes	5.4583	1.22899	6
		Total	5.3448	1.72095	29
	Total	No	5.1484	1.73072	32
		Yes	5.5833	1.84116	15
		Total	5.2872	1.7585	47

10.2 - Second Pilot Study - Test of Within-Subject Effects							
Source		**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intrinsic	Greenhouse-Geisser	5.138	1.676	3.066	2.913	0.07	0.063
*Intrinsic * English*	Greenhouse-Geisser	4.568	1.676	2.725	2.589	0.091	0.057
*Intrinsic * Philosophy*	Greenhouse-Geisser	3.211	1.676	1.915	1.82	0.175	0.041
*Intrinsic * English * Philosophy*	Greenhouse-Geisser	0.218	1.676	0.13	0.124	0.849	0.003
Error (Intrinsic)	Greenhouse-Geisser	75.862	72.074	1.053			

10.3 - Second Pilot Study - Test of Between-Subject Effects						
Source	**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intercept	1136.857	1	1136.857	447.087	0.000*	0.912
English	0.048	1	0.048	0.019	0.891	0
Philosophy	0.036	1	0.036	0.014	0.905	0
*English * Philosophy*	1.158	1	1.158	0.456	0.503	0.01
Error	109.341	43	2.543			

Table 11.

Second Pilot – Extraneous Load

11.1 - Second Pilot Study - Descriptive Statistics					
Lecture	**English**	**Philosophy**	**Mean**	**Standard Deviation**	**N**
Piketty (P1)	No	No	4.8056	0.86402	9
		Yes	3.9444	2.84983	9
		Total	4.375	2.09033	18
	Yes	No	2.9022	2.05567	23
		Yes	0.6667	0.6455	6
		Total	2.4397	2.0601	29
	Total	No	3.4375	1.98685	32
		Yes	2.6333	2.74816	15
		Total	3.1809	2.25886	47
Rousseau (R2)	No	No	4.25	1.05327	9
		Yes	3.25	2.9128	9
		Total	3.75	2.18619	18
	Yes	No	3.1304	1.93189	23
		Yes	1.25	0.80623	6
		Total	2.7414	1.91036	29
	Total	No	3.4453	1.78788	32
		Yes	2.45	2.47162	15
		Total	3.1277	2.05751	47
Tocqueville (T3)	No	No	3.8611	1.51096	9
		Yes	3	2.33854	9
		Total	3.4306	1.96065	18
	Yes	No	2.8804	1.9712	23
		Yes	1.2917	0.8279	6
		Total	2.5517	1.89852	29
	Total	No	3.1563	1.88345	32
		Yes	2.3167	2.02984	15
		Total	2.8883	1.94964	47

11.2 - Second Pilot Study - Test of Within-Subject Effects							
Source		**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Extraneous	Sphericity Assumed	1.976	2	0.988	1.581	0.212	0.035
*Extraneous * English*	Sphericity Assumed	8.208	2	4.104	6.567	0.002*	0.132
*Extraneous * Philosophy*	Sphericity Assumed	0.501	2	0.251	0.401	0.671	0.009
*Extraneous * English * Philosophy*	Sphericity Assumed	0.529	2	0.264	0.423	0.657	0.01
Error (Intrinsic)	Sphericity Assumed	53.745	86	0.625			

11.3 - Second Pilot Study - Test of Between-Subject Effects						
Source	**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intercept	319.001	1	319.001	96.762	0.000*	0.692
English	31.037	1	31.037	9.414	0.004*	0.18
Philosophy	18.249	1	18.249	5.535	0.023*	0.114
*English * Philosophy*	2.286	1	2.286	0.693	0.41	0.016
Error	141.761	43	3.297			

Table 12.

Second Pilot – Comprehension

12.1 - Second Pilot Study - Descriptive Statistics (Absolute Score)					
Lecture	**English**	**Philosophy**	**Mean**	**Standard Deviation**	**N**
Piketty (P1)	No	No	5.8	1.874	10
		Yes	5.67	2.598	9
		Total	5.74	2.182	19
	Yes	No	6.3	2.055	23
		Yes	5.5	2.168	6
		Total	6.14	2.065	29
	Total	No	6.15	1.986	33
		Yes	5.6	2.354	15
		Total	5.98	2.099	48
Rousseau (R2)	No	No	5.4	1.647	10
		Yes	6	1.871	9
		Total	5.68	1.734	19
	Yes	No	4.91	2.021	23
		Yes	6.33	2.733	6
		Total	5.21	2.21	29
	Total	No	5.06	1.903	33
		Yes	6.13	2.167	15
		Total	5.4	2.029	48
Tocqueville (T3)	No	No	6.9	1.663	10
		Yes	5.78	1.787	9
		Total	6.37	1.77	19
	Yes	No	6.13	2.096	23
		Yes	6.67	2.503	6
		Total	6.24	2.149	29
	Total	No	6.36	1.981	33
		Yes	6.13	2.066	15
		Total	6.29	1.989	48

12.2 - Second Pilot Study - Test of Within-Subject Effects							
Source		**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Comprehension	Sphericity Assumed	10.483	2	5.241	2.454	0.092	0.053
*Comprehension * English*	Sphericity Assumed	0.288	2	0.144	0.067	0.935	0.002
*Comprehension * Philosophy*	Sphericity Assumed	12.396	2	6.198	2.902	0.06	0.062
*Comprehension * English * Philosophy*	Sphericity Assumed	6.61	1.91	3.461	1.547	0.22	0.034
Error (Intrinsic)	Sphericity Assumed	187.98	88	2.136			

12.3 - Second Pilot Study - Test of Between-Subject Effects						
Source	**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intercept	1344.354	1	1344.354	477.098	0.000*	0.916
English	0.024	1	0.024	0.009	0.926	0
Philosophy	0.065	1	0.065	0.023	0.88	0.001
*English * Philosophy*	0.862	1	0.862	0.306	0.583	0.007
Error	123.982	44	2.818			

Table 13.

Second Pilot – Recoded Comprehension

13.1 - Second Pilot Study - Descriptive Statistics (Percentage)					
Lecture	**English**	**Philosophy**	**Mean**	**Standard Deviation**	**N**
Piketty (P1)	No	No	0.48	0.18738	10
		Yes	0.4778	0.25386	9
		Total	0.4789	0.21494	19
	Yes	No	0.5217	0.2044	23
		Yes	0.4833	0.24833	6
		Total	0.5138	0.20997	29
	Total	No	0.5091	0.19743	33
		Yes	0.48	0.2426	15
		Total	0.5	0.21037	48
Rousseau (R2)	No	No	0.5	0.18257	10
		Yes	0.5667	0.15	9
		Total	0.5316	0.16684	19
	Yes	No	0.4435	0.21068	23
		Yes	0.5833	0.25626	6
		Total	0.4724	0.22344	29
	Total	No	0.4606	0.20146	33
		Yes	0.5733	0.19074	15
		Total	0.4958	0.20312	48
Tocqueville (T3)	No	No	0.5545	0.13853	10
		Yes	0.4545	0.17604	9
		Total	0.5072	0.16125	19
	Yes	No	0.4822	0.18678	23
		Yes	0.5455	0.22268	6
		Total	0.4953	0.19221	29
	Total	No	0.5041	0.17471	33
		Yes	0.4909	0.19376	15
		Total	0.5	0.17889	48

13.2 - Second Pilot Study - Test of Within-Subject Effects							
Source		**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Recoded Comprehension	Sphericity Assumed	0.02	2	0.01	0.469	0.627	0.011
*Recoded Comprehension * English*	Sphericity Assumed	0.009	2	0.005	0.215	0.807	0.005
*Recoded Comprehension * Philosophy*	Sphericity Assumed	0.095	2	0.048	2.19	0.118	0.047
*Recoded Comprehension * English * Philosophy*	Sphericity Assumed	0.047	2	0.024	1.089	0.341	0.024
Error (Intrinsic)	Sphericity Assumed	1.913	88	0.022			

13.3 - Second Pilot Study - Test of Between-Subject Effects						
Source	**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Significance**	**Partial Eta Squared**
Intercept	9.792	1	9.792	383.85	0.000*	0.897
English	0	1	0	0.007	0.934	0
Philosophy	0.004	1	0.004	0.172	0.68	0.004
*English * Philosophy*	0.011	1	0.011	0.415	0.523	0.009
Error	1.122	44	0.026			