Pre-publication version of Santiago, J., Meseguer, P., & Valenzuela, J. (2025). Reaction time. En: Rojo López, A. M., & Muñoz Martín, R. (Eds.). Research Methods in Cognitive Translation and Interpreting Studies. (pp. 324-346). John Benjamins Publishing Co. https://doi.org/10.1075/rmal.10.14san

Chapter 14

Reaction time

Julio Santiago (1), Purificación Meseguer (2) & Javier Valenzuela (2)

1 University of Granada | 2 University of Murcia

Reaction time is the time it takes for a person to respond to a stimulus. Reaction time has proven useful in assessing aspects such as processing speed or coordination and is widely used in disciplines such as cognitive psychology or neuroscience. This chapter gives an overview of the history of the method and covers its main contributions to different fields, with a special focus on CTIS. It will also address conceptual, ethical, and methodological issues, in an attempt to provide the reader with the necessary keys to understand its potential and applications in research.

Reaction time refers to measuring the interval between a stimulus presentation and the participant's response. As a research method, it yields valuable insights into the temporal dynamics of cognitive and motor processes, making it foundational in experimental psychology and related disciplines. This chapter introduces the method, delves into its historical context, and explores its significant contributions across various fields, emphasizing its relevance in CTIS. It also discusses conceptual and practical considerations, including key methodological and ethical aspects, to equip readers with a comprehensive understanding of its research applications and emerging challenges.

1. The method and key questions

Reaction times are defined as the time span between the presentation of a stimulus and the response to that stimulus. Reaction time is a relevant measure of psychological functioning given the central assumption that the time it takes to go from a stimulus to a response reflects the complexity of the processing that the nervous system carries out during that time. Thus, when reaction time is measured in specific experimental designs, it is often possible to infer sequences of processing stages as well as the nature of the processing within each stage. Reaction timing is thus a valuable method to reveal the structure and workings of the mind that underly the skill under investigation. And, as we shall see in

the following pages, it has proved useful in CTIS, providing interesting insights into the processes of translation and interpretation.

1.1 The state of the art in reaction time research

The interest in reaction time started in 1820 when F. W. Bessel, an astronomer from Könisberg, compared the speed at which he was able to respond to the passing of a star through the hairline of his telescope with that of a visitor to his observatory. Despite these early attempts, in physiology and psychology the speed of the nervous impulse and of thought itself was generally considered infinite and unmeasurable until 1850, when Helmholtz first designed an experimental set up to measure the speed of nervous transmission in a frog (Helmholtz, 1850, cited in Schmidgen, 2002). This rendered a surprising discovery: nerve impulses progressed at the relatively slow speed of about 30 m/s. Helmholtz proceeded to find out very similar speeds in humans using ingenious procedures, and he noted that nerve conduction took only a small portion of total reaction time. Central processing took the lion's share and became the main interest of later reaction time studies (Brebner & Welford, 1980). In 1868, Donders developed an intuitive subtractive logic method that allowed the drawing of substantive inferences about mental processes occurring between stimulus and response. 1.1.1 The discovery of processing stages Simple reaction time is the time it takes to respond to a stimulus in experimental tasks when there is only one possible stimulus and one possible response. That is, the person expects a specific stimulus and knows exactly what response should be produced when the stimulus is detected. However, the person does not know exactly when the stimulus will arrive. In this situation, Donders (1868) established that the times that participants need to respond are about 200–250 ms (with some differences between sensory modalities). During that interval, several physiological and psychological processes unfold and peripheric nerve transmission only accounted for a few milliseconds. How could the duration of central processes be measured? Donders reasoned that, in a simple reaction time task, the reaction interval is occupied by the detection of the stimulus and the release of the command to execute the pre-formed response. If, instead, the situation allowed for several possible signals, each one cueing the performance of a different action (a procedure called a choice reaction time), the reaction interval should be increased by the time needed to discriminate the signal and choose among the response alternatives (see Figure 1).

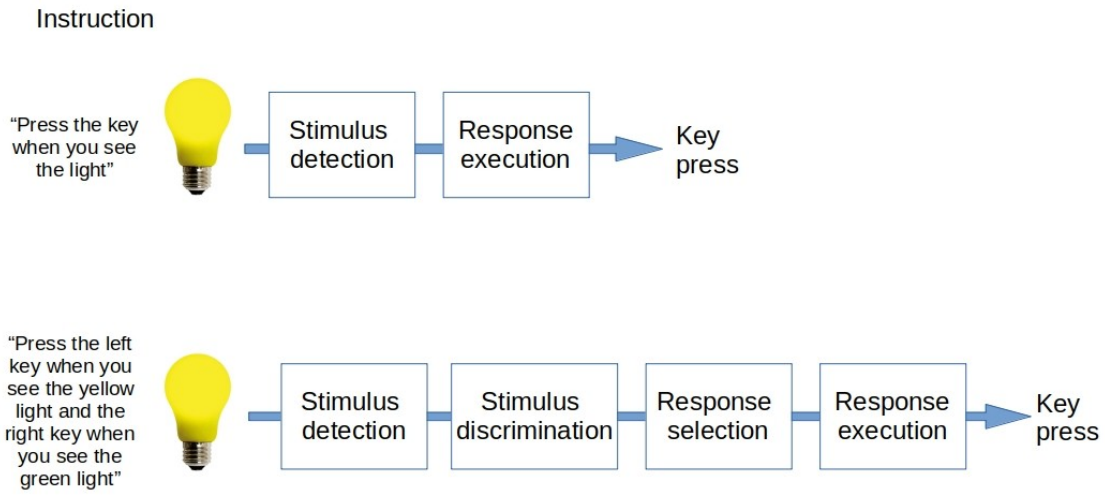

Figure 1. Processing stages in simple (above) and choice (below) reaction time procedures

Donders (1868) tested this prediction in different setups, where he varied the modality of stimuli (visual, auditory, tactile) and responses (producing a fixed syllable, pronouncing the presented syllable), as well as several procedural features (e.g., number of response alternatives). In all cases, choice reaction times were longer than simple reaction times by between 60 and 200 ms. Elusive mental processes such as decision making took a measurable amount of time. He extended his subtractive logic even further, by using a procedure that later became known as a go-no go task. In this experimental paradigm, participants may receive several signals, but they only need to respond to one of them, always in the same way. Here, the reaction time would be devoted to stimulus detection and discrimination and also to deciding whether to respond, but the response is already preformed, selected, and ready to be released. In other words, the stages of response selection and assembly are removed from the interval. Donders found that, under specific conditions, the choice task increased the reaction time 83 ms over simple reaction time, whereas the go-no go task increased reaction time only 36 ms. He therefore estimated the duration of the response selection and assembly stages in 47 ms. By combining experimental designs with the subtractive method, it was possible to analyze the time between stimulus and response into discrete processing stages.

Donders' (1868) work introduced reaction time in the toolbox of cognitive psychologists. The popularity of this method went up and then down during the first decades of the 20th century, along with the interest in studying unobservable mental processes. It also started a specific way of thinking about reaction times, based on the idea that the reaction interval comprises several, self-contained processing stages and that stage durations could be measured by comparing tasks with different processing structures. This brought about new problems for researchers: how could they be certain that a given change in a task would add only the hypothesized additional stage without altering the other stages? Moreover, when a task was modelled to consist of certain stages, how could they validate that model and contrast it with alternative analyses?

With the cognitive revolution of the 60s, the use of reaction times gathered momentum again. Sternberg (1969) proposed a significant elaboration of Donders' method. The subtractive method focused on comparisons between tasks with or without specific processing stages. In contrast, Sternberg (1969) assessed the effects of different factors on a single task. A factor is an experimental treatment with two or more levels (e.g., responding to a dim versus a bright light). The effect is the difference in reaction time between levels. The logic of additive factors follows from the notion of processing stages as modular, sequential components. A stage is completed after its component processes finish and their results are passed on to the next stage, which in turn does its work or, if final, releases the response. In this view, a factor that prolongs a stage (say, the brightness of the stimulus increases the time needed for the stimulus discrimination stage) should add a fixed amount to total reaction time. If another factor affects a different stage (say, having to choose between possible responses increases the time needed for response selection), it should add another fixed amount to total reaction time, and both amounts should add up to yield the total increase in reaction time. Additive effects of this kind appear in charts of experimental results as characteristic patterns of parallel lines (see Figure 2). In comparison, if two factors affect the same stage, their effects will result in a multiplicative effect, such that the increase due to one factor will be modulated by the other factor. Therefore, the finding of additive versus multiplicative effects of different factors can be used to diagnose whether they are affecting the same or different processing stages.

Let us use one of Sternberg's (1969) experiments to illustrate the additive factor method. In this experiment, participants had to name aloud single digits presented visually on a screen. Reaction time was measured from the presentation of the digit to the beginning of the vocal response. Sternberg manipulated three factors in a way that allowed him to assess the effects of all possible combinations of their levels onto reaction time. The first factor was stimulus quality: the digit could be presented with standard clarity or degraded by superimposing a checkerboard pattern. The second factor was response compatibility: the compatible response was naming aloud the presented digit; a less compatible response was naming aloud the number resulting from adding one to the presented digit. The third factor was the number of options: in one condition there were only two possible digits, either 1 or 8; in the other condition all eight digits from 1 until 8 could be presented.

If two factors are simultaneously manipulated in a task and their effects are additive, this would suggest that the factors affect independent processing stages.

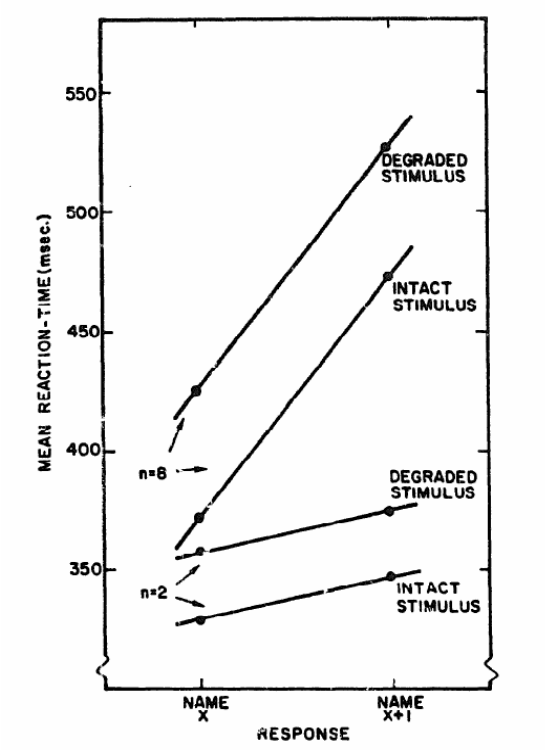

Figure 2. Results of experiment 5 in Sternberg (1969). The effect of stimulus degradation is additive with the effects of response compatibility and number of alternatives, whereas number of alternatives interacts with both stimulus degradation and response compatibility

As shown in Figure 2, the interpretation of results revealed that stimulus quality and response compatibility had additive effects: degrading the stimulus lengthened reaction time by the same amount when the response was compatible and when it was less compatible, and vice versa. In contrast, the number of response alternatives (two vs. eight) affected reaction time more strongly when the stimulus was degraded; and it also had a stronger effect when the response was less compatible. This pattern of results suggests that the task is composed by, at least, two processing stages: a stage of stimulus encoding (affected by stimulus degradation, but not by response compatibility) and a stage of response selection (affected by response compatibility, but not by stimulus degradation). The number of alternatives affected both stages by having more stimuli to discriminate and more responses to select from.

Sternberg's additive factors method makes it possible to draw inferences about the existence and properties of the processes carried out during these stages. However, it also has limitations: it does not allow estimations of the duration of stages, and it does not hint at the order of stages. It is also limited to sequential arrangements of processes, where each stage must finish before the next one starts. This leaves out overlapping processes. The additive factor logic also does not apply to cases where all the processes in the chain draw resources from a limited capacity pool such that the total sum of demands affect reaction time (Welford, 1980a). Finally, this method does not allow a bottom-up induction of the processing architecture from the data. In other words, interpreting the pattern of combined effects on

reaction times calls for a theory of internal processes, i.e., a theory of the processing stages involved and how they combine. This theory can then be used to derive hypotheses about which factors are likely to affect any processing stage and what the resulting patterns may be of additivity and interaction.

By the end of the 60s, the basic conceptual tools that allowed the use of reaction times to study the architecture of cognition were in place: First, mental processes took a measurable amount of time. Second, they were organized in complex architectures; some processes were organized sequentially, other in parallel. Third, theories about processing models (including component processes, nature of their computations, their independence and interdependence, and serial vs. parallel organization) could be tested with careful experimental designs, including cross-task comparisons and the crossed manipulation of factors.

1.1.2 Some basic findings in the fields of language and memory

By the 1960s and 70s, a sizeable empirical database of basic findings had been compiled, and these findings revealed details of the workings of unobservable mental processes (see Welford, 1980b). On the perceptual side, it was well established that reaction times were affected by stimulus intensity and quality, and by sensory modality — tactile signals were responded to faster than acoustic signals and these, in turn, faster than visual signals (e.g., Donders, 1868; Sternberg, 1969). It was also known that decision time increases linearly with the number of bits of information needed to encode the number of alternatives (Hick, 1952). Memory search increases linearly with the number of items to be searched (Sternberg, 1969). Responses using more effectors, requiring greater coordination among effectors, or composed of a longer series of motor elements increase the latency to start the response, although only up to a point, suggesting that people can also carry out motor planning processes during the execution of longer or more complex responses (Klapp & Wyatt, 1976). Reaction time is reduced when stimuli are presented at expected locations or at the location of signals which automatically attract attention, even when the eyes have not moved to those locations, showing the deployment of attention (Posner, 1980). Mental rotation of images was carried out at a constant speed, as shown by the linear relation between degrees of rotation and reaction time (Cooper & Shepard, 1973). Virtually all areas of research within cognitive psychology have benefitted from using reaction times and have continued doing so. In this section we mention just a few findings that are especially relevant to language and memory processing, fields which, among others, are of special interest to students of the cognitive processes of translation.

Frequency effects. Finding a frequency effect of a given linguistic unit on its reaction time (and other measures) has customarily been considered evidence for the existence of a stored mental representation of that unit, on the rationale that a frequently reused unit representation will end up being permanently stored in memory. Early reports observed that lexical frequency affected the latency of discriminating words from nonwords (pronounceable sequences of letters or phonemes that do not constitute words in the language). In this lexical decision task, frequent words are discriminated from nonwords faster than less frequent words (Scarborough et al., 1977). Many models of language processing assume that there are stored lexical unit representations (often conceived as nodes in a network) and frequency effects

arise because high-frequency unit representations have a lower activation threshold or a higher resting activation level (e.g., Dell, 1986). Similarly, the storage of other linguistic unit representations, such as syllables, has been proposed based on the finding of frequency effects on lexical decision and naming tasks (Álvarez et al., 2004).

Priming effects. When the processing of a stimulus of certain characteristics (the target) is preceded by the processing of another stimulus with which it holds some relation (the prime), the time taken to respond to the target may change as compared to when it is preceded by a control (another stimulus unrelated to the prime and/or the target). This is called a priming effect, which is interpreted as evidence of the psychological reality of the relation that holds between the prime and the target. Priming effects have been extensively used to study the structure of connections between memory representations.

The first study of a priming effect, Meyer and Schvaneveldt (1971), focused on processing semantically related words. They presented participants with two letter strings on a screen and they had to judge whether one of them (in one experiment) or both of them (in another experiment) were words vs. nonwords. The conditions of interest were those in which both strings were words (thus, the response was kept constant). In the priming condition, the two words had a semantic relation (e.g., bread-butter), whereas in the control condition there was no such relation (bread-doctor). Latencies were shorter in the priming than in the control condition by an averaged 101 ms across both experiments. This effect suggests that accessing the meaning of a word is affected by the retrieval of a semantically related word. Meyer and Schvaneveldt (1971) already used this effect to argue for some models of lexical memory over others. Overall, most subsequent research is consistent with the idea that the mental lexicon is a network of nodes linked to each other along meaning-related lines. Under this view, activating the node of a prime word sends activation through its links to other nodes, such that when the target word is presented, its node is already pre-activated and therefore takes a shorter time to be recognized. Much debate remains, though, with the only point of agreement being that the priming effect is a useful research paradigm.

Since this original study, the most common priming designs present primes and targets sequentially and require responding only to the target stimuli. When word processing is of interest, the most commonly used tasks have been lexical decision and overt naming of the target. However, priming designs have been modified in many ways to address specific questions. The following section reviews some of the research questions and findings explored in CTIS.

1.1.3 Some reaction time findings in CTIS

Reaction time has been mostly used in CTIS to explore issues of bilingual lexical processing in written translation, focusing on the processing of words out of context rather than on translation at the textual level. Nevertheless, research results are still valuable: for example, central questions tapped by these studies are whether words are always mapped onto concepts, or whether translation is carried out on the basis of a common conceptual representation or a word-toword association at the lexical level (Brysbaert et al., 2014; García, 2015). Despite some disagreement on the issue of how language information is organized at each level of representation (shared or separately for each language), it is

generally agreed that there is a common conceptual (or semantic) store where the meaning of words and sentences is organized (see Francis, 2005, for a review). We know that some of the most easily translated words (that is those with shortest RTs) are those that are cognates (words with morphophonological resemblance in a given pair of languages; e.g. intelligent in English vs. inteligente in Spanish; de Groot, 1992a), concrete words, highly frequent words, or words high in imageability (words which evoke a clear mental image, e.g., car; de Groot, 1992b; de Groot et al., 1994). Other findings from reaction time experiments have also shown that the direction of translation is relevant, as it is faster to translate into the L1 than into the L2 (Chou et al., 2021).

Reaction time has also been used to explore the role of political ideology. For example, Rojo and Ramos (2014) designed a priming study to explore the influence of translators' political and economic beliefs (i.e., left-wing/liberal vs. right-wing/ conservative) on the time they needed to find a translation equivalent. Their results showed that the type of prime exerted a significant influence on the time participants took to find a suitable translation: Words that were incongruent with participants' ideological viewpoint caused longer reaction times than words that were congruent with their ideological beliefs. The authors also found differences between the effect of primes on each group: Whereas left-wingers were faster when reading a word consistent with their beliefs, right-wingers showed no difference in reaction time between the two conditions.

In a later study on a similar topic, Rojo and Meseguer (2021) tested whether translation students' position on Catalonia's bid for independence (in favor, against, and neutral) influenced the time they took to read newspaper headlines on Catalonia's independence crisis and choose between three potential translation equivalents: literal, in favor, and against. Participants' reaction times were expected to be influenced by the (in)congruence between the expressions and the participants' position in the conflict, but results from the study reported no significant interaction for the congruency between participants' ideology and the content of source text or translation options. However, the effect of participants' ideology when reading-to-translate the headlines was statistically significant, with participants against independence being slower than those in favor and those being neutral. It seemed that participants against Catalonia's independence had greater difficulties when reading-to-translate headlines on a topic that per se conflicted with their ideology.

Some studies have also used reaction time as an instrument to measure interpreters' behavior. Three noteworthy studies that focused on conference interpreters are the works by Chmiel (2016, 2018, 2020). The first study compared the performance of unidirectional vs. bidirectional interpreters to analyze the influence of directionality on the speed of lexical retrieval. Results showed a directionality effect across all participants, with interpreters providing faster L2-L1 translations. However, when comparing groups, a directionality effect was found only for bidirectional interpreters, which could mean that the predominant direction of interpreters is not the only factor that influences the asymmetry of lexical and conceptual links in the mental lexicon. Chmiel also reveals in her study an expected context effect, with high contextual constraint shortening translation latencies in bidirectional interpreters too: when reading high contextual constraint sentences, interpreters analyze them semantically, anticipating the sentence-final word and providing the translation faster.

In the second study, Chmiel (2018) proposed a priming study to explore the influence of expertise and training on word recognition and cross-linguistics connections in the mental lexicon comparing, this

time, the performance of professional interpreters and interpreter trainees. Results revealed faster word recognition in advanced trainees when compared to beginners, but slower when compared to professionals, whereas the priming effect was just found in the L1-L2 direction. The author claimed that interpreter training could impact the speed at which interpreters recognize words.

In the third study, Chmiel (2020) explored once again the effect of training and experience on word-translation accuracy, latency, and anticipation. Professional conference interpreters and interpreter trainees at the beginning and at the end of their two-year training program were first exposed to an incomplete sentence and then asked to interpret the final word when appearing on screen. They completed the task from and to their native language. Results showed that anticipation was not enhanced by either training or experience, but it could be modulated by the direction of translation, as the anticipation effect was stronger in the L1-L2 direction. Results also showed that word-translation latency was decreased only by training, without a significant effect of professional experience. In contrast, professional interpreting experience facilitated inhibition and the selection of an appropriate translation equivalent.

To sum up, reaction time stands out as a promising method in CTIS, although its application to translation and interpreting tasks still faces important methodological challenges, some of which will be addressed in the next section.

1.2 Ethical issues

Reaction times are a non-invasive method that does not pose specific ethical concerns. Reaction time tasks are usually simple tasks such as pressing a key in response to a tone or pronouncing a presented word aloud. The high amount of random noise in reaction times (see discussion below) often leads to the use of many pairs stimulus-response (trials), which may lengthen the task and make it feel repetitive and boring to the participant, but without posing any risk to their integrity. Moreover, as boredom and repetition lead to inattention and this, in turn, to even greater variability in latencies, experimenters try and keep task duration at very reasonable levels, usually well below one hour. Thus, only general ethical issues must be taken into consideration, such as securing that participation in the study is voluntary, that participants are well informed about who are the responsible researchers and how to contact them, what is the task to be done, that they can stop the task and leave at any moment, that they can retire their consent to use their data for the study, and that they give explicit consent to participate. Participants should also be informed about the treatment that will be given to their data and provide their consent. Usually, data will be anonymized and reported only in aggregate form, what relieves concerns regarding data protection.

Of course, some studies may bring up particular ethical concerns, for example, when sensible materials such as erotic or violent texts are used or when the study targets special populations such as minors or clinical samples. There may also be local legal regulations to take into account. This is why all studies must always obtain approval by the local Ethics Review Board before starting data collection.

2. Conceptual aspects

2.1 Reaction time as dependent variable

Since the first reports, time measurements smaller than one thousandth of a second (a millisecond) were deemed psychologically insignificant. This follows from the relatively low speed of nervous transmission first measured by Helmholtz (Schmidgen, 2002). Milliseconds are, still, fleeting time units that accumulate very fast upon the slightest distraction of the participant or inaccuracy of the measuring apparatus. High-precision clocks are widely available nowadays, but the accurate measurement of reaction time also depends on the precise detection of the instant of stimulus presentation and the response onset, and at both sides there are technical challenges that introduce both random and systematic noise. Furthermore, many factors may affect reaction time, lengthening or shortening latencies by a few or a great many milliseconds. As a result, reaction time is a highly variable measure. It allows researchers to detect subtle influences on behavior but always within a sea of noise due to unrelated causes. Reducing the noise, both through specific procedures and through statistical analyses (see Section 2.2.), becomes an essential part of using reaction time in empirical studies.

Statistically, reaction times present the added challenge of not following a normal distribution. For example, a normal latency to name a printed word aloud under ideal conditions may amount to about 500 ms (half a second), with times smaller than 250 ms being usually considered the result of anticipations. Any slight difficulty experienced during the reaction interval may push this latency up for a great amount. Giving a second look to an infrequent word, glancing away from the location of stimulus presentation for a moment, and many other causes may make the participant need an extra second before they are able to respond, what amounts to adding 1000 ms to the average latency, four times the minimum possible (non-anticipatory) latency.

A more common distraction may push up the latency over the average by several thousands of milliseconds. As a result, reaction time distributions usually approximate the shape of a normal distribution over the range of shorter latencies, but with a long rightward tail of longer but less frequent latencies (Figure 3). Most of this overall shape (tail included) can be considered to result from processes that mediate between stimulus and response, but the tail usually contains some latencies which are clearly not the result of the processes of interest (e.g., a fortuitous distraction). All in all, reaction times, by their own nature, require that the researcher adopts strategies to (1) minimize and control noise; (2) remove responses that are not caused by the processes of interest; and (3) either bring the distribution into the shape of a normal distribution or use analyses that can be applied to other distributional patterns.

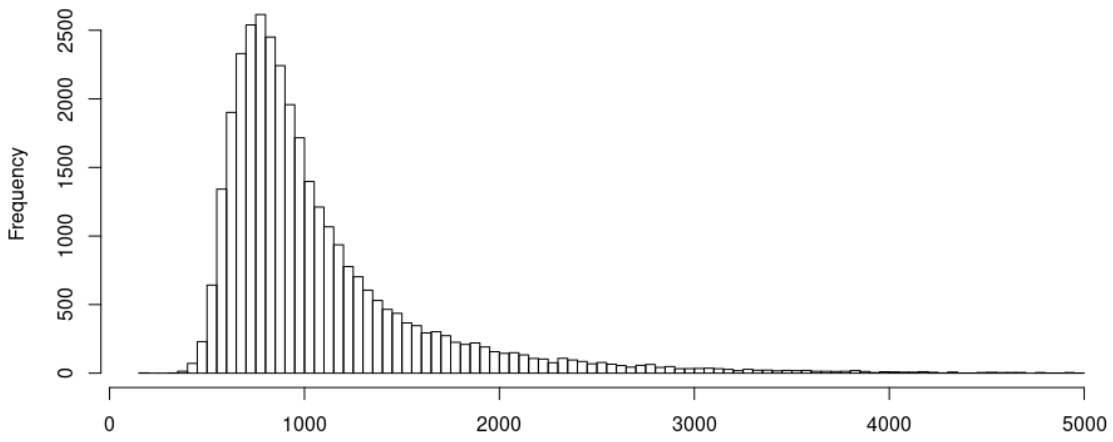

Figure 3. A typical reaction time distribution

2.2 Controlling noise in reaction time research

Noise is the variation in the data that is not due to factors of interest for the researcher. Most — if not all — behavioral measures suffer from an important level of noise. As discussed above, reaction times are a particularly noisy variable. Some noise is systematic in the sense that it is due to the effect of uncontrolled factors, and reaction times can be affected by very many causes (see below for a taxonomy of noise sources in reaction time designs). There is also an unknown amount of purely random noise. Noise makes it more difficult to detect the effect of the factor of interest among all the clutter, but noise may pose an even more perverse problem. Sometimes uncontrolled factors may influence latencies in a way that mislead the researcher to think that the factor of interest is having an effect while it is not. Noise control refers to both minimizing the amount of random noise (which results in an increase of the precision of measurement) as well as to making sure that there are not confounding influences in the data. Controlling noise is an essential part of all experimental designs.

Basically, noise can be controlled by reducing the amount of unaccounted variation. If researchers find out what the uncontrolled factors are and are able to control them, they can reduce the amount of unaccounted noise. Very importantly, researchers must take measures to ensure that the noise is equally distributed over the conditions being compared. In other words, whatever the factors that affect one side of a comparison, they should also affect the other side of the comparison to the same extent (see examples below). Three groups of factors can introduce systematic noise in a reaction time experiment, related to the participants, the materials, and the experimental setup:

1. Participant-related factors are individual factors that may affect the speed of participants' responses (e.g., attentiveness, motivation, familiarity with computers, prior practice, and so on).

2. Material-related factors are factors characteristic of the items with which participants interact and respond to in the study. With simple stimuli and responses, they include sensory modality (visual, auditory, etc.), stimulus characteristics (brightness, clarity, etc.) and the type of motor responses (key presses, foot movements, etc.). CTIS research projects often use complex stimuli such as words and

sentences, which also vary in many factors that may affect reaction time (e.g., frequency, length, syntactic, morphological and phonological structures, and so on). In tasks such as picture naming, imagerelated factors (e.g., picture complexity) are important as well. In tasks involving translation, two linguistic systems come into play, which may increase the number and kinds of potential material-related factors.

3. Factors related to the experimental setup are those related to the conditions in which testing is carried out, including the equipment (e.g., sensitivity of the microphone, screen refresh rate), but also room temperature, lighting, environmental noise, distracting stimuli, and the like.

Noise control must be approached in view of the question of interest. Very few, if any, reaction time studies are concerned with the absolute value of the latency to respond in a given condition. Instead, the studies are concerned with the difference in reaction time between two or more conditions, defined by the levels of one or more factors. Crucially, the comparison of interest, together with the design used, defines the potential sources of systematic noise.

Systematic noise can be controlled basically in two ways: (1) by equaling the conditions being compared on the factor producing the noise; and (2) by making the noise due to that factor to be distributed randomly over the two sides of the comparison. Take, for example, a study aimed at finding out whether very frequent words are translated faster than less frequent words. In such a study, a set of highly frequent and a set of less frequent words are selected and both sets are presented to all participants. Technically, word frequency is a fixed factor, because its levels are established (fixed) by the researcher. Here, the contrast of interest is the time it takes to respond to high vs. low frequency words. People differ considerably in their speed of responding (e.g., because of their age, experience with computers, motivation, etc.).

However, all participants see both sets of words, so there is no reason to worry about any participant-related factor affecting the contrast. This is due to two reasons. First, both sides of the contrast (high vs. low frequency words) will be equally affected by this participant-related systematic noise. Because the contrast of interest (high vs. low frequency words) compares a sample of participants with themselves (as they respond to both sets of items), data collected on both sides of the comparison will be equaled on the effect of all and every participantrelated factor, even those unknown to the researcher. Therefore, there is no risk that any participant-related factor misleads the researcher to falsely believe that high and low frequency words have different latencies if they do not. Second, statistical procedures will take each participant as a level of a factor in the design (call it Participant) and will attribute all systematic noise linked to individual differences to this factor. Technically, Participant is a random factor, because its levels are assumed to be extracted randomly from the population of reference (this also has important consequences for generalization, as discussed below). Thus, statistical analyses will quantify all participant-related noise precisely, and take it away from the random noise term in the analyses, thereby reducing the amount of unaccounted noise. In other words, participant-related noise will not contribute to the clutter and the effect of the fixed factor will be detected more easily over the remaining noise.

In a design like this one, noise can be cancelled equally effectively for both known and unknown sources of participant-related noise because we are making use of a within contrast, one that occurs

within a given unit of analysis. In this case, the unit is the participant, and the contrast is within-participants. Withincontrasts are ideal regarding noise control, but they are only effective regarding the factors that introduce systematic noise over that unit. In the example above a more difficult problem of noise control arises when we focus on the items (the words) as the unit of analysis. In this case, the study will compare a set of highly frequent words with a set of infrequent words. This comparison is a between contrast because it occurs between different levels of a unit of analysis.

A between contrast will be plagued by systematic noise introduced by the myriad factors, both known and unknown to the researcher, on which that unit of analysis (words in the example) differ. Some of those factors may actually correlate with the factor of interest, potentially introducing confounds. For example, less frequent words tend to be longer than highly frequent words, and word length by itself may increase response latencies. Thus, uncontrolled word length may actually produce a difference in reaction time between high and low frequency words, leading to wrong conclusions in the study. Additionally, uncontrolled factors that do not correlate with the factor of interest will not pose the risk of misleading the researcher as to the cause of the effect, but they will introduce systematic noise that cannot be disentangled from purely random noise over the contrast by statistical means, thereby making it more difficult to detect the effect of the fixed factor.

Contrasts between groups of a given unit of analysis (usually either participants or items) are therefore affected by unit-related factors that introduce systematic noise. In this case, the researcher should strive to identify those factors and actively match the conditions of the contrast in them. We might endeavor to match the high and low frequency words in their length, as we should also do with many other factors known to affect word processing. However, there will always be unknown factors which we will not be able to act upon. The only strategy available to control for unknown factors in a between contrast is to select the units randomly from the population of all units having the characteristics of interest (in this example, the sets of high and low frequency words in the language). Random sampling of sizeable unit sets will achieve that unknown factors will affect both sides of the comparison equally. They will introduce noise but, at least, it will not be confounding noise. Regarding noise control, between contrasts necessarily pose a more difficult problem than within contrasts. Therefore, the researcher should try to avoid between contrasts and test within contrasts whenever possible.

Sometimes turning a between contrast into a within contrast is possible by changing the design. For example, instead of comparing words which are already of high versus low frequency in the language, we could artificially increase or decrease the frequency of each word by presenting it very often or not at all to the participants before they do the experimental task. We can then compare how participants behave with high versus low frequency words while having the same words at both sides of the contrast. To do so, we might start the word frequency experiment by selecting a set of words of intermediate frequency, and divide them randomly into two halves. We then choose a set of texts (call it set A) where one half of those words appear very often while the other half never occurs, and another set of texts (set B) where the former half of words never occurs while the latter half of words occurs very often. Then we ask two groups of randomly selected participants to read the texts, such that each group reads only one set. In doing so, we will have increased the frequency of occurrence of each word in one group and decrease it in the other group, effectively turning the comparison between high and low frequency words into a within-contrast. Unfortunately, this kind of strategy is not always feasible

(e.g., when we want to compare speakers of different languages in a given task), and we will have to content ourselves with controlling for as many known sources of noise as possible over the relevant unit of analysis (participants, words, pictures…) and randomly choosing from all those units that meet the selection criteria.

Random choice of the elements of a given unit of analysis (i.e., a random factor) is important because it facilitates noise control, but there is another reason why it is crucial: because it affects our ability to generalize the findings. If the levels of the random factor (e.g., the participants) are randomly extracted from the population of possible levels (all possible participants), then the results of the study regarding the effect of the fixed factor can be generalized to the population defined by the random factor. The units of analysis over which language studies usually define their contrasts are both the participants and the linguistic items, and researchers will be interested in generalizing the findings both to the population of participants as well as to the population of linguistic items.

As for the third source of systematic noise, the experimental setup, the best advice is to keep setup-related noise to a minimum by using technologies specially designed for reaction time experiments (such as experiment generators and voice keys) and to take measures to secure that sources of setup-related noise affect equally all participants and items. In this way, most contrasts of interest will be within-setup, and therefore this source of noise will be controlled for.

3. Implementation

3.1 Data pre-processing in reaction time research

In reaction time tasks, erroneous responses are usually considered to arise from processes differing from those that generate correct responses in substantial ways. Moreover, distributions of the reaction times of correct responses are rightskewed (Figure 3) for reasons discussed above. Although part of this rightskewness is due to the processes of interest, part of it is due to the intromission of unrelated causes (e.g., momentary absent-mindedness due to tiredness), usually causing long latencies. Additionally, on some occasions participants may anticipate their response and produce by chance a correct response with a very short latency. Therefore, part of the standard analysis of reaction times starts with a stage of data separation and trimming.

First of all, data from correct trials are separated from data from error trials. Latencies from error trials are usually not analyzed any further, but the number of errors is compared over the contrasts of interest. In other words, latencies and accuracy are often analyzed in parallel and, in fact, it is necessary to do so because of the possibility of speed-accuracy trade-offs. Participants may strategically decide to emphasize speed over accuracy when responding to a particular condition of the task, or they may emphasize accuracy over speed. The former leads to shorter latencies and more errors, and the latter to longer latencies and fewer errors, but for reasons unrelated to the factor of interest.

In order to keep these trade-offs constant over the conditions, the instructions usually ask participants to be both fast and accurate. If the task generates enough number of errors, it is possible to check for the presence of speedaccuracy trade-offs in the data by comparing latency and accuracy results. Usually, latency and number of errors will be affected in the same direction by the fixed factor: more

difficult conditions should lead to both longer latencies and more errors (and vice versa). Longer latencies in conditions with a smaller number of errors (or shorter latencies with a greater number of errors) suggest a potential underlying trade-off. If there are very few errors or there are no significant effects in the analysis of accuracy, trade-offs with speed are of no concern. When there are trade-offs it is necessary to use some statistical methods that are able to combine the information provided by latency and accuracy into a single measure (Liesefeld & Janczyk, 2019).

Latencies from correct trials must then be trimmed, firstly, with the goal of removing outlier latencies, because they are the result of irrelevant processes. Secondly, if the statistical analysis is going to use parametric methods (see the next section), trimming is also used to bring the overall distribution closer to a normal distribution, as it is required by those methods. There are several procedures for this. One simple trimming procedure is to apply cut-offs, thresholds below and above which the data are removed from the analysis (either hard cut-offs or cutoffs based on the individual mean and standard deviation). Another possibility is to transform the data (e.g., using logarithmic or inverse transformations). Reaction time distributions are usually still highly right-skewed after cutting off the extremes, so it is common to use both cut-offs and transformations. Each trimming and transformation option has advantages and disadvantages (see Ratcliff, 1993). However, note that non-linear transformations and central tendency measures such as medians and geometric means cannot be used when the goal is applying the additive factors method to analyze processing stages (Sternberg, 1969), as they distort the additive-interactive patterns between factors.

3.2 Statistical analysis in reaction time research

Data preparation for analysis is highly dependent on what kind of statistical analysis is intended. By far, the most popular statistical analysis of reaction time data (and many other measures used in experimental designs) is Analysis of Variance (ANOVA). One important problem of ANOVA is that it assumes that the data are normally distributed and latencies hardly ever comply with this requirement. As discussed above, if ANOVA (or any other parametric analysis) is going to be used, the data are trimmed with the goal to bring them closer to normality.

ANOVA breaks the total amount of variation in the data into two parts: one that is accounted for by the experimental conditions (the levels of the fixed factor), and one that is unaccounted noise. ANOVA calculates the proportion between the variation due to the fixed factor and the unaccounted noise, and estimates the probability that the former has been observed by chance. If this probability is low (usually below 0.05 or 5%), researchers can claim that the finding is significant (in this case, that the fixed factor is probably producing a real effect on the latencies). When the contrast between the experimental conditions is a within-contrast (a comparison over a unit of analysis, usually participants and items, what we have called a random factor), ANOVA also estimates which part of the unaccounted noise is due to variation over the unit of analysis, thereby removing it from the noise and decreasing the amount of unaccounted noise which increases the sensitivity of the analysis to detect the effect of the fixed factor.

An important problem with ANOVA is that it can only take a single random factor into account. As discussed above, language studies often include two random factors, participants and items, and it is

important to be able to generalize to both populations, participants and items. In the past, this was solved by computing independent ANOVAs by participants and by items, or by combining both into a single statistic (Clark, 1973). Recent developments in statistics have brought forward a different analytical solution which is nowadays widely considered to be superior to ANOVA: linear mixed models (introduction in Baayen, Davidson, & Bates, 2008; best practice guide in Meteyard & Davies, 2020). Linear mixed models cater for two (or more) random factors simultaneously. Besides, mixed models are not constrained to work on normally distributed data. Generalized mixed models can be used with many different kinds of distributions. Specifically, regarding reaction times, Lo and Andrews (2015) recommend to trim the data to remove outliers but avoid applying any transformations. Instead, two analyses of the data should be carried out, one assuming a Gamma distribution and another assuming an Inverse Gaussian distribution. Both distributions reproduce the surface characteristics of reaction time distributions (basically, the right skewness). The analysis yielding a better fit to the data is then used to assess the contrasts of interest.

As it may have become obvious to the reader, researchers have many degrees of freedom at their disposal when analyzing reaction times and there are many choices regarding how to preprocess and analyze the data. However, this flexibility may have negative consequences on the quality of inferences to be drawn from the data. As mentioned, keeping the probability of asserting a significant effect under 5% when there really is none is a central aspect of statistical inference and, therefore, of the trust on scientific findings: it maintains the probability of false alarms at a level that is accepted as low enough. Yet, flexibility in the analysis may raise the probability of false alarms well above 5% (Simmons, Nelson, & Simonsohn, 2011). By trying different kinds of trimming, normalization, and analysis strategies, and then selectively reporting only those strategies that produced significant findings, the reaction time researcher may easily rise the probability of false alarms well over 17% (Morís & Vadillo, 2020). Together with other common practices in scientific research and publishing (e.g., p-hacking, file drawer problem, and, chiefly, the use of low-powered studies), Ioannidis (2005) claims that more than 50% of all published studies in biobehavioral sciences are false alarms. Researchers should, therefore, apply strict countermeasures to avoid them (good practice guide in Munafò et al., 2017).

3.3 Practical recommendations

At this stage, we would like to offer some practical recommendations before introducing the relevant software applications. Generally speaking, it is advisable to choose a design and procedure that aligns with the research questions. Clear hypotheses and an analysis plan should be established, linking each hypothesis to a specific statistical test. Ethical guidelines should be adhered to and approval obtained from the local Ethics Committee. Researchers should strive for control for participant-, item-, and setup-level confounds using equalization, counterbalancing, and/or randomization, with within-contrasts when possible as well as plan exclusion criteria for participants and outliers. It is also advisable to run a power analysis before data collection, specifying expected effect size, power goal, and necessary sample size for detection. A detailed analysis plan covering data collection, outlier detection, data trimming, and statistical analysis should be developed, using appropriate tests for the specific design and research question. Ideally, hypotheses, exclusion criteria, analysis plan, statistical

power, and sample size analysis should be pre-registered.

A familiarization phase should be included in the task, in order to reduce the likelihood of losing initial trials. It should also be verified that the chosen program presents the correct number of trials in each condition, and the proper functionality of counterbalances and randomizations confirmed. It is advisable to ensure well-installed equipment, checking battery levels for portable devices; no study-irrelevant software should be running in the background during data collection. All raw and processed data should be securely backed up. Also, causal data interpretation should be approached with caution: the limitations of the design chosen should be explicitly stated, and current best practices guidelines in reporting followed. The materials, programs, (anonymized) raw data, and analysis scripts should be deposited in a public repository for easy re-analysis, meta-analyses, and experiment replication. Finally, it should be ensured that shared programs and scripts are comprehensible, functional, and allow detailed replication of reported findings.

These are the main take-home reminders. Let us now consider the software applications.

3.3 Software for reaction time research

Researchers who are interested in unravelling mediators' behavior will find reaction time a very useful method, as it allows to calculate with a very high level of precision the time that elapses from the time the subject receives a stimulus until the subject provides a response to that stimulus which, as was mentioned in previous section, could offer invaluable information about the stages and phases of the translation process. There is currently a plethora of software for measuring reaction times on the market; Bridges et al. (2020) conducted a study (which we strongly recommend reading before embarking on a reaction time experiment) in which they compared the timing performance of several behavioral science software packages, on various operating systems, and in both laboratory-based "native" systems and on studies conducted remotely via web-browser. The different packets were tested in Windows, MacOS and Ubuntu. In spite of the complex casuistry, due to the multiple combination of OS and browsers, apparently, for lab-based studies, PsychoPy, Psychtoolbox, NBS Presentation and E-Prime turned out to be the most precise, closely following by OpenSesame and Experyment. In on-line packages, PsychoPy was the best performer, with a precision of under 4 ms in every browser/OS combination. Regarding operative systems, it tended to depend on the type of study (e.g., for visual stimuli, MacOs seems to lag behind), but overall, Windows seems to be the most stable performer across all domains.

4. Closing remarks

4.1 Emerging challenges in reaction time research

Reaction time is a useful method for understanding the intricacies of the mind. This explains its allure to researchers interested in cognitive science and language, including, of course, CTIS researchers. Reaction time might uncover many aspects of translation or interpreting processing which so far remain largely unexplored. Nevertheless, using reaction time to study translation and interpreting is

complicated. Translation is a cluster of complex processes that involve the activation of two different linguistic systems and a process of converting one into the other, and the results of this activity must be evaluated considering a great number of contextual demands. In contrast with such a potential complexity, reaction time studies seem better suited for simple tasks: those that require quick responses to very concrete stimuli. This could be one of the reasons why most of the studies related to translation and interpreting have tended to be focused on tasks such as the translation or interpreting of isolated words (Chmiel, 2018; Rojo & Ramos, 2014) or on reading for translation (Macizo & Bajo, 2004, 2006). Generally speaking, reaction time studies in translation are better applied to translation tasks which are carefully narrowed down to some of its subprocesses, such as the choice of strategies. However, over and above the difficulty of the enterprise, the relative scarcity of works which use this method in translation research could perhaps be related to the non-existence of a tradition in this area: the field needs to accrue a "critical mass" of researchers using this methodology so that it becomes part of the repertoire of research methods.

Nonetheless, there are still many phenomena and processes in translation — such as comprehension, attention, or memory — that can be investigated and can benefit from reaction time experiments. The results of the research carried out so far and the wide possibilities open to researchers are a clear invitation to make use of this method, which no doubt will prove its worth in the quest for a more precise understanding of the processes involved in translation and interpreting.

Further readings on reaction time

Baayen, R. H., & Milin, P. (2010). Analyzing Reaction Times. International Journal of Psychological Research, 3(2), 12–28.

Jensen, A. R. (2006). Clocking the mind: Mental chronometry and individual differences (1st ed.). Elsevier.

Welford, A. T. (Ed.). (1980). Reaction times. Academic Press.
https://en.wikipedia.org/wiki/Mental_chronometry

References

Álvarez, C. J., Carreiras, M., & Perea, M. (2004). Are syllables phonological units in visual word recognition? Language and Cognitive Processes, 19(3), 427–452.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390–412.

Brebner, J. M. T., & Welford, A. T. (1980). Introduction: An historical background sketch. In A. T. Welford (Ed.), Reaction times (pp. 1–23). Academic Press.

Bridges, D., Pitiot, A., MacAskill, M. R., & Westley Peirce, J. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. PeerJ 8: e9414.

Brysbaert, M., Ameel, E., & Storms, G. (2014). Bilingual semantic memory: A new hypothesis.

Foundations of Bilingual Memory. 133–146.

Chmiel, A. (2016). Directionality and context effects in word translation tasks performed by conference interpreters. Poznań Studies in Contemporary Linguistics, 52(2), 269–295.

Chmiel, A. (2018). Meaning and words in the conference interpreter's mind: Effects of interpreter training and experience in a semantic priming study. Translation, Cognition & Behavior, 1(1), 21–41.

Chmiel, A. (2020). Effects of simultaneous interpreting experience and training on anticipation, as measured by word-translation latencies. Interpreting. International Journal of Research and Practice in Interpreting, 23(1), 18–44.

Chou, I., Liu, K., & Zhao, N. (2021). Effects of directionality on interpreting performance: evidence from interpreting between Chinese and English by trainee interpreters. Frontiers in Psychology, 12, 781610.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior, 12(4), 335–359.

Cooper, L. A., & Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), Visual information processing (pp. 75–176). Academic Press.

De Groot, A. M. B. (1992a). Bilingual lexical representation: A closer look at conceptual representations. In R. Frost & L. Katz (Eds.), Orthography, phonology, morphology, and meaning (pp. 389–412). Elsevier.

De Groot, A. M. B. (1992b). Determinants of word translation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 1001–1018.

De Groot, A. M. B., Dannenburg, L., & van Hell, J. G. (1994). Forward and backward word translation by bilinguals. Journal of Memory and Language, 33(5), 600–629.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. Psychological Review, 93(3), 283–321.

Donders, F. C. (1868). On the speed of mental processes. Acta Psychologica, 30, 412–431.

Francis, W. S. (2005). Bilingual semantic and conceptual representation. In J. F. Kroll & A. M. B. de Groot (Eds.), Handbook of bilingualism: Psycholinguistic approaches (pp. 251–267). Oxford University Press.

García, A. (2015). Psycholinguistic explorations of lexical translation equivalents: Thirty years of research and their implications for cognitive translatology. Translation Studies, 4, 9–28.

Hick, W. E. (1952). On the rate of gain of information. Quarterly Journal of Experimental Psychology, 4(1), 11–26.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLOS Medicine, 2(8), e124.

Klapp, S. T., & Wyatt, E. P. (1976). Motor programming within a sequence of responses. Journal of Motor Behavior, 8(1), 19–26.

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs. Behavior Research Methods, 51(1), 40–60.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. Frontiers in Psychology, 6.

Macizo, P., & Bajo, T. (2004). When translation makes the difference: Sentence processing in reading and translation. Psicológica, 25, 181–205.

Macizo, P., & Bajo, T. (2006). Reading for repetition and reading for translation: Do they involve the same processes? Cognition, 99, 1–34.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. Journal of Memory and Language, 112, 104092.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. Journal of Experimental Psychology, 90(2), 227–234.

Morís Fernández, L., & Vadillo, M. A. (2020). Flexibility in reaction time analysis: Many roads to a false positive? Royal Society Open Science, 7(2), 190831.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie, N., Simonsohn, U., & Wagenmakers, E. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1(0021).

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716-1-8.

Posner, M. I. (1980). Orienting of attention. Quarterly Journal of Experimental Psychology, 32, 3–25.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. Psychological Bulletin, 114(3), 510–532.

Rojo López, A. M., & Ramos Caro, M. (2014). The impact of translators' ideology on the translation process: A reaction time experiment. MonTI, Special Issue 1, 247–271.

Rojo López, A. M. & Meseguer Cutillas, P. (2021). The effect of attitude towards Catalonia's independence on response latency when translating ideologically conflicting press headlines. Onomazein Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile, 8, 128–145.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. Journal of Experimental Psychology: Human Perception and Performance, 3(1), 1–17.

Schmidgen, H. (2002). Of frogs and men: The origins of psychophysiological time experiments, 1850–1865. Endeavour, 26(4), 142–148.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11), 1359–1366.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. Acta Psychologica, 30, 276–315.

Welford, A. T. (1980a). Choice reaction times: Basic concepts. In A. T. Welford (Ed.), Reaction times (pp. 73–128). Academic Press.

Welford, A. T. (Ed.). (1980b). Reaction times. Academic Press.