Behavior Research Methods (2022) 54:1976–1988
https://doi.org/10.3758/s13428-021-01707-1

Generalizability of pause times in sentence production to distinguish between adult writers

Catherine Meulemans[1] · Sven De Maeyer[2] · Mariëlle Leijten[1]

Accepted: 11 September 2021 / Published online: 23 November 2021

Abstract

Researchers often decide on the number of trials included in an experiment without adhering to an empirical method or framework. This might compromise generalizability and unnecessarily increase participant burden. In this article we want to put forward generalizability theory as a guide for task reduction. We will use a sentence production task to demonstrate how a generalizability and a decision study can help researchers to estimate the minimum number of trials and of items per trial that are necessary to generalize over trials. We obtained writing process data for 116 participants. Each of them completed a sentence production task that had 40 trials. Pause times between and within all words, target nouns and target verbs were logged with the keystroke logging tool ScriptLog. Results demonstrate that generalizability theory can serve as an empirical framework to ensure generalizable measurements on the one hand, and reduce participant burden to a minimum on the other. This finding is particularly valuable for studies with vulnerable target groups, such as participants suffering from aphasia, dyslexia or Alzheimer's disease.

Keywords Generalizability theory · Sentence production · Task reduction · Writing processes · Keystroke logging

Introduction

When setting up an experiment, a researcher is faced with a number of decisions. One of them concerns determining the number of trials that each participant is asked to complete. This is an important choice because, as classical test theory (CTT) states, the number of trials influences the reliability of the measurements (Cronbach et al., 1972). CTT refers to reliability as the degree to which a measurement score reflects a participant's true score. It assumes that a true score would be obtained if a measurement did not contain any errors. However, some error is always present in a measurement, which implies that not the true score, but an observed score is obtained. This observed score can therefore be split into a true score and a random error term. Reliability is then defined as the proportion of true score variance to observed score variance (Bloch & Norman, 2012; Cronbach et al., 1972).

✉ Catherine Meulemans
catherine.meulemans@uantwerpen.be

[1] Department of Management, University of Antwerp, Antwerp, Belgium

[2] Department of Training and Education Sciences, University of Antwerp, Antwerp, Belgium

Reliability gives insight into the number of trials one should carry out. If you use a test with 100 multiple-choice questions, you will get a better idea of your knowledge on the subject that is tested than when the test contains only one multiple-choice question. By increasing the number of observations, or in this case the number of questions, the error variance will decrease. Hence, the observed score variance will decrease as well, leading to a higher proportion of true to observed score variance, or reliability (Cronbach et al., 1972). The same holds for scientific experiments: the more trials participants complete, the more reliable the measurements should be.

With this in mind, it would be quite logical to maximize the number of trials if researchers want to ensure that measurements are sufficiently reliable. However, long and extensive experiments impose more burden on participants than short and efficient ones, which makes it undesirable to include a very large number of trials (Bradburn, 1978; Ulrich et al., 2005). When participant burden is high, participants' motivation and concentration drop, leading to distorted results (Allen et al., 2016; American Psychological Association, 2002; Cohen et al., 2017; Graham et al., 2018; Troia et al., 2013). Moreover, higher perceived participant burden is associated with lower likelihood of participation in an experiment (Lingler et al., 2014).

Generalizability theory to balance reliability and participant burden

Designing a scientific experiment is a balancing act between protecting the well-being of those who participate and ensuring the reliability of the collected measurements. To achieve this balance, generalizability (G) theory can be used as a framework (Brennan, 1992, 2001; Cronbach et al., 1972; Shavelson et al., 1989; Shavelson & Webb, 1991). G theory is a method to evaluate measurement reliability and is considered a more comprehensive alternative to CTT (Mushquash & O'Connor, 2006).

In contrast to CTT, G theory allows researchers to split measurement error into more than one source. These sources of error are called "facets" and may be any characteristic feature of the measurement, such as task, trial, item, occasion or rater. Each facet contains multiple levels, such as the different tasks (task A and B) or the different measurement occasions (1st and 2nd moment), which are called "conditions". When every condition within a facet co-occurs with each condition of another facet, facets are crossed. In contrast, when facets are nested, each condition within a facet only appears with one condition of another facet (Bloch & Norman, 2012; Mushquash & O'Connor, 2006). G theory allows researchers to unravel the impact of all facets on measurements, and to identify how they are linked to one another (i.e., being nested or crossed).

Scholars who have published on G theory (e.g., Shavelson & Webb, 2006) propose a two-part analysis, in which first a generalizability (G) study is carried out to evaluate the extent to which these facets contribute to measurement inaccuracy. The facets are identified, followed by an estimation of the variances in measurement caused by each of these facets and their interactions. Based on these variance estimations the generalizability score of the observed measurement is described. This score or "G coefficient" indicates to what extent the measurements can, for instance, be generalized from one trial to another (Shavelson & Webb, 2006). In the second part of the analysis, the results of the G study can be used to perform a decision (D) study in which alternative scenarios are explored. The error variances of the G study are used to estimate how changing the number of observations in one or more facets influences measurement generalizability (Mushquash & O'Connor, 2006; Shavelson & Webb, 2006). This allows researchers to tackle questions such as: How many trials do we need to generalize over trials? What is the lowest number of items each trial should include? And how do we find the right balance between the number of trials and the burden imposed on participants? The aim of the current paper is to show how this application of G theory can support researchers in estimating the ideal number of trials and items per trial. We will illustrate this with an example taken from the field of writing research. To make this illustration more concrete, a step-by-step walkthrough of the analyses is made available at https://gifted-nightingale-d45b7e.netlify.app. However, we would first like to give some background information on writing research and the application of G theory in this field.

G theory in previous writing research

Written language production has mainly been approached from two perspectives: (1) writing single words (e.g., Bertram et al., 2015; Purcell et al., 2011; Weingarten et al. 2004, and (2) composing full texts in various fields of study, such as educational research (e.g., Bouwer et al.,2018; De Smedt et al., 2016; Fidalgo et al., 2015), second and foreign language research (e.g., Cheong et al., 2019; Knospe et al., 2019; van Weijen et al., 2009), and developmental research (e.g., Kellogg, 2008; Mateos & Solé, 2009). The application of G theory in studies of writing tasks focusing on full texts has already been explored, allowing researchers to successfully determine the ideal number of writing assignments both within and across text genres (Bouwer et al., 2015; Gebril, 2009; Graham et al., 2016; Schoonen, 2005; van den Bergh et al., 2012).

However, another way to study written language production is with the use of sentence production tasks (Ford & Holmes, 1978; Ronald T. Kellogg, 2004; Nottbusch, 2010; Nottbusch et al., 2007). A written sentence production task is a task in which a specific sentence is elicited (often with images or text). As with single word production, its written output is largely driven by the input stimuli and can, therefore, be kept constant across trials. Since output characteristics such as sentence structure and word length partly determine the underlying writing process, this will be more stable if its resulting product remains constant as well. Hence, sentence production tasks allow for more accurate writing process comparisons thanks to their controlled output (Nottbusch, 2010). This control makes sentence production tasks suitable for observing the planning of words, phrases and syntactic structures in writing.

When researchers want to study, for instance, planning in writing, they do so mainly by looking at how long and how often someone pauses before and while writing the unit of interest (e.g., word, phrase, sentence or text) (Leijten et al., 2019; Medimorec & Risko, 2017; Wengelin, 2006). A pause can be defined as a moment of non-writing that lasts longer than the motoric transition between two keys. However, pause times do not only reflect planning, but can also occur during revision and sometimes even translation. This is because, in contrast to motor execution, pauses offer a space for these demanding writing processes to take place (Alves et al., 2008). Hence, observing pauses is important to fully grasp what is happening when someone writes, because it gives

insight into the cognitive processes that take place during the course of writing (Wengelin, 2006).

However, pauses exist only temporarily during the writing process, which makes them difficult to study. To capture the writing process and make pauses tangible, keystroke logging is used (Sullivan & Lindgren, 2006; Wengelin, 2006; Wengelin et al., 2009). Keystroke logging is a technique to register the entire writing process by recording each keystroke and the time that elapses before, after and during its use (Leijten & Van Waes, 2013). Each writing process recording therefore contains measurements on the actual typing as well as on the pauses in-between, making those pauses analysable. Nottbusch (2010), for instance, studied pauses in a sentence production task to look at planning of two syntactic structures: coordinated versus subordinated subject noun phrases. The former include two noun phrases that are syntactically equal (e.g., the brown squirrel and the red strawberry), whereas in the latter, two noun phrases are included, the second of which is subordinate to the first one (e.g., the brown squirrel with the red strawberry). Results showed that the subordinated structure required more cognitive effort and was therefore planned before production onset. This was reflected by the longer pause times that were recorded before the start of typing. In contrast, pause times were more evenly distributed in the coordinated structure, indicating a rather gradual planning of its production.

To determine the number of trials, Nottbusch (2010) used the number that was necessary to have all items occur eight times in each of the possible positions in the sentence. In another study, Nottbusch et al. (2007) asked participants to type 24 sentences twice: first based on a picture and then by copying those same sentences from a written presentation. Similarly to Nottbusch (2010), the number of sentences was chosen specifically so that the six stimuli items could each be elicited in a coordinate and subordinate sentence structure but also to ensure that the last, predicative noun of the sentence was written an equal number of times in singular and in plural. In both cases, the researchers used the number that was necessary to have all items occur in all different conditions. However, there also exist other pragmatic approaches for determining the number of trials. Kellogg (2004), for example, instructed students to write sentences based on two prompt nouns. In this case, Kellogg (2004) based the number of items per trial on an earlier study that he replicated and in which two nouns were used as well (Power, 1985). Yet, in numerous studies the number of sentences and filler sentences are reported as a given without further justification (Leijten et al., 2011; Negro et al., 2005; Quinlan et al., 2012).

As illustrated above, the number of trials included in a sentence production task is frequently decided upon without the support of an empirical method or framework. In those instances, researchers' decisions are rooted in practical or even intuitive reasons. This pragmatism could potentially compromise measurement generalizability, increase participant burden and, in the worst case, render the experiment useless. Whether in the aforementioned studies these approaches led to measurements that were generalizable enough to draw conclusions from, remains unclear. However, in comparison to pragmatic methods of task reduction, G theory could have been applied as a framework to guarantee measurement generalizability.

The aim of the current paper is to show how G theory can assist researchers in determining the length of an experiment while preserving the balance between generalizable measurements and participant burden. We will illustrate this by estimating the absolute minimum number of trials and of items per trial in a sentence production task that would still allow researchers to reliably draw conclusions about differences between participants regarding between- and within-word pause times.

Method

Participants

The data analysed in the present paper were gathered from 116 volunteers who were recruited through associations for elderly people and members of the personal network of the main researcher. Participants' ages varied between 50 and 90 years, with 35 participants aged 50–59 years, 35 aged 60–69 years, and 46 above 70 years of age. All participants had Dutch as their mother tongue and were accustomed to typing on a standard Belgian (azerty) keyboard.

These data were collected in an exploratory study that was part of a larger research project on the writing process characteristics of healthy aging elderly and Alzheimer's patients. The aim of this exploratory phase was twofold: (1) to explore the evolution of writing processes in healthy ageing, and (2) to inform and prepare follow-up of cross-sectional and longitudinal studies by examining how many trials were needed in the experiments. The study was reviewed and approved by the Ethics Committee of Antwerp University Hospital (Committee for Medical Ethics; Belgian reference number B300201629701).

Materials, procedure and design

Participants were asked to complete a sentence production task on a computer. The task had 40 trials, in which participants needed to type a sentence based on one, two or three images (from the Open Linguistic Picture Database; Paesen & Meulemans, 2020) and a verb. Every image depicted a single object that had been chosen to evoke a predefined noun. The verbs were elicited by presenting participants with the verbs' infinitive forms written in full. As shown in Fig. 1, these input

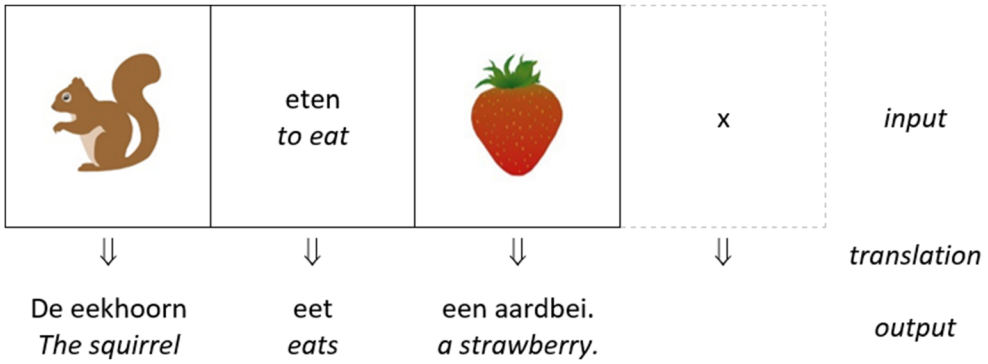


Fig. 1 Example trial of a sentence production task

stimuli needed to be horizontally translated into a sentence. This implies that participants had to name the images as correctly as possible, conjugate the verbs, and then incorporate these elements into a sentence. The task was administered in Dutch.

The sentence production task was designed to elicit three sentence structures of different length (see Fig. 2). The intransitives were assigned 20 trials. The same was applied to the transitives (divided into mono- and ditransitive in Fig. 2), with an equal number being allocated to the mono- and ditransitive structures (i.e., 10 trials each)

Thus, the trials with intransitive sentences each contained three words and those with mono- and ditransitive sentences contained five and eight words, respectively. This resulted in an average of 4.75 words per sentence. Moreover, the trials with intransitive sentences each contained one target noun, whereas those with mono- and ditransitive sentences contained two and three nouns, respectively. This amounts to 1.75 nouns per sentence on average, whereas the number of verbs per sentence equalled 1 regardless of the sentence structure.

The sentence production task consisted of a fixed set of images and verbs that were pseudorandomly combined into trials. This resulted in a design in which images plus verbs equalled words (*w*), whereas trials equalled sentences (*s*) composed from those words. Each combination of trials was then assigned to a participant (*p*). In this design, these words, sentences and participants were facets (i.e., sources of measurement error). Each participant received the same images and verbs (i.e., words) but distributed differently across the 40 trials. Since the distribution differed per participant, the sentences themselves were unique. Hence, each condition of the facet "words" co-occurred with each condition of the facet "sentences", whereas each condition of the facet "sentences" only co-occurred with one condition of the facet "participants". This created a G study design in which words were crossed with sentences, and sentences were nested within participants (p:sw).

Data collection and variables

The task was administered on a laptop and logged with ScriptLog, a keystroke logging tool that registers the entire writing process (Frid et al., 2014). Keystroke logging makes a writing process tangible by recording each pressed keystroke and the time that elapses before, after and during its use

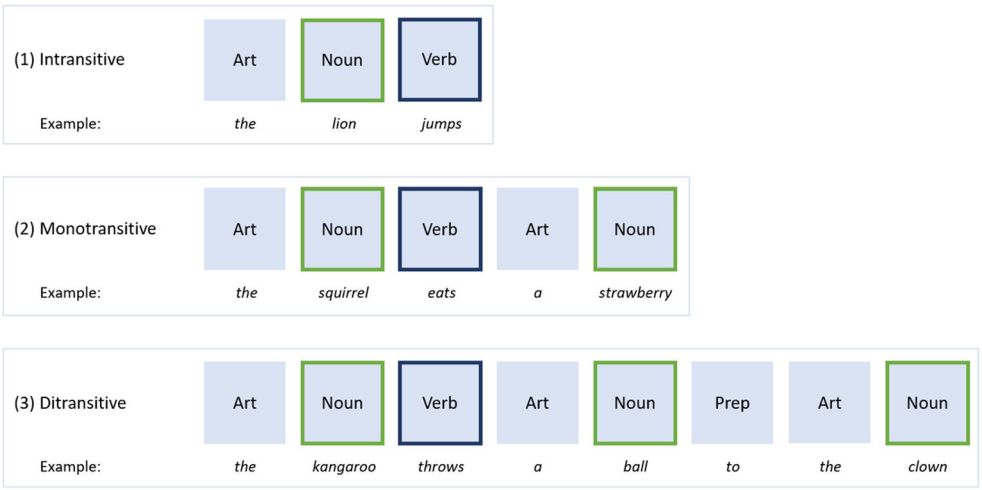


Fig. 2 Overview of sentence structures with indication of target word categories (nouns and verbs) per structure. *Note*. Art = article; Prep = preposition

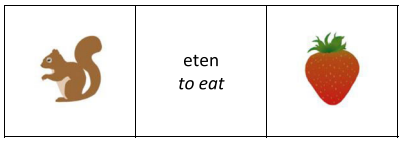
(Leijten & Van Waes, 2013). The result can be considered a detailed representation of the actions that take place from blank sheet to finished text.

Because pause times will be the primary focus of the cross-sectional and longitudinal follow-up studies, we also focused on these pause variables for the G and D studies. For the calculation of these pause times the key in-key in principle was used. This means that the calculation for pauses is based on the time that elapses between key presses (as opposed to key releases). Pause data on a total of 4,640 sentence production processes were collected. We eliminated the trials during which participants made deletions and corrections, since we were only interested in pause times related to fluently written sentences. The final data set contained words from the 3,180 remaining sentences (68.53%), which equalled 14,544 words, or an average of 27 sentences per participant.

Pauses can occur at different moments during the writing process: before the participants start to type a sentence, between two subsequent words, and even between single characters of the same word. Those pause locations can be defined more clearly by specifying the words they occur with. In this study, pauses can occur in combination with words of all kinds of categories (i.e., all words) or before specific target nouns or target verbs (i.e., target words). Therefore, pauses between and within all words, target nouns and target verbs will form the six focus variables of this paper. When we refer to pause times between target nouns or pause times between target verbs, we refer to all pauses that occur between the preceding word and the noun or verb that follows (i.e., after a random word and before the target). Figure 3 shows an example of how the focus variables were analysed: to examine the pause time between all words, we would look at the interruptions before the first characters of the words "squirrel", "eats", "a" and "strawberry". However, to study pause times within target nouns in this example, we would look at the moments of non-activity between the characters of "squirrel" and "strawberry".

Analysis

The statistical tool selected for the analyses was the R software (R Core Team, 2019) and more specifically the package *lme4*(Bates et al., 2015). To gain more insight into the generalizability of the pause times, we first conducted a number of G studies. Each focused on one variable, resulting in a total of six studies (e.g., pause times between and within all words, nouns and verbs). In all of them, the sentences and words were potential sources of error, also known as facets, because word characteristics (e.g., word length, word frequency, age of acquisition) and specific (e.g., unexpected versus likely) combinations of words when put into sentences influence the length of pauses (Brysbaert et al., 2014; Medimorec & Risko, 2017; Scaltritti et al., 2016). The aim was to quantify the amount of variance (S^2) in each variable that is caused by these facets, both for each facet separately and regarding their interaction.



De eekhoorn eet een aardbei.
The squirrel eats a strawberry.

between-word pause time

within-word pause time

# character	output	pause time	pause location	
1	T	3722	before sentence	
2	h	520	within word	
3	e	104	within word	
4	*SPACE*	*96*	*after word*	**1472 ms = between-word for *squirrel***
5	**s**	**1376**	**before word**	
6	q	1288	within word	
7	u	528	within word	
8	i	233	within word	
9	r	847	within word	
10	r	232	within word	
11	e	240	within word	
12	l	208	within word	
13	*SPACE*	*160*	*after word*	
14	e	352	before word	
15	a	229	within word	**216 ms = within-word for *t***
16	**t**	**216**	**within word**	
17	s	232	within word	
18	*SPACE*	*168*	*after word*	
19	a	336	before word	
20	*SPACE*	*176*	*after word*	
21	s	408	before word	
22	t	296	within word	

...

Fig. 3 Overview of pause times

As mentioned before, words were crossed with sentences, which made it essential to estimate the interaction of these two facets as well. Consequently, the variance in pause times is the sum of the following variances:

$$S^2_{pause\ time} = S^2_{participant} + S^2_{sentence} + S^2_{word} + S^2_{sentence \times word} \quad (1)$$

To know the extent to which each of these facets contributed to the variance in pause times, we ran a linear mixed model (function *lmer()* of package *lme4*) in R that estimated the variance proportions associated with each of the facets by modelling the facets as random effects. Personal typing speed (based on a copy task) potentially affected pause times and was, therefore, also included in the linear mixed model as a fixed effect. This way, the variance estimates in the model represent variances for participants with an average typing speed. The model can be written as

$$\gamma_{wsp} = \beta_0 + \beta_1 \times Copy\ task_p + \left(\nu_{0p} + \mu_{0ps} + \psi_{0w} + \epsilon_{0psw}\right) \quad (2)$$

with γ_{wsp} being the score of a pause time γ for a certain word w, in a sentence s written by participant p. Pause times were expected to depend on the facets: p(participant) reflected in the participant-specific deviation towards the intercept ν_{0p}; s(sentence) reflected in the sentence-specific deviation towards the intercept μ_{0ps}; and w(word) reflected in the word-specific deviation towards the intercept ψ_{0w}. All interactions between these facets were captured in the residual term ϵ_{0psw}.

Subsequently, the variance estimates based on the random effects and the number of sentences ($N_{sentence}$) and words (N_{word}) were entered in Formula 3 to estimate the generalizability of differences in pause times between participants. This led to a generalizability ($E\rho^2$) score for the pause time associated with the estimated variances for the facets. This score can be interpreted as the generalizability of differences between participants over the sentences and words written by the participants. Given that the generalizability coefficient can be interpreted as a reliability measure, acceptable generalizability was set at .70 or higher. This means that 70% of the variance observed between participants is due to the participants themselves, while the other 30% is due to the other facets. This cut-off is also suggested by van den Bergh et al. (2012) and Bloch and Norman (2012).

$$E\rho 2 = \frac{S^2_{participant}}{S^2_{participant} + \frac{S^2_{sentence}}{N_{sentence}} + \frac{S^2_{word}}{N_{word}} + \frac{S^2_{sentence \times word}}{(N_{sentence} \times N_{word})}} \quad (3)$$

Using the results of each G study, a decision (D) study was performed. Each of these D studies can be considered a simulation that explores how the generalizability of a variable (i.e., pause times) is influenced by a change in the number of observations in each of the facets. In the exemplary sentence production task, it gives insight into how a different number of sentences or words per sentence influences a variable's generalizability over participants. Based on these results, the number and length of the trials that are necessary to reliably estimate the pause times is decided upon.

A step-by-step walkthrough of the R script that was used for these analyses and the generation of the plots can be found at https://gifted-nightingale-d45b7e.netlify.app. Data can be downloaded separately from https://osf.io/h2fcw/download.

Results

First, the generalizability of the current sentence production task was estimated. The variance proportions of the following components were estimated: participants, sentences, words, and the interaction of words and sentences (i.e., random error). As shown in Fig. 4, the variance proportions differed depending on the estimated pause times. Participants alone determined a small part of the total variance (1.71%) in the pause times that occur between all words, whereas they were responsible for over 12% of the variance in pause times within all words. The opposite occurred for the facet "words": this component was responsible for over 68% of the variance of the pause times between words, in comparison with close to 8% for pause times within all words. Target words followed the same pattern for the variance caused by participants (nouns: 7.31% and 16.02% for between- and within-word pause times, respectively; verbs: 4.16% and 19.19% for between- and within-word pause times, respectively). For the word component the variance tended to be higher within than between target words, which is the opposite of what was observed for all words. However, the variances attributed to the word component never reached 8% for any of the pause times related to target nouns or verbs. Finally, sentences determined less than 1% of the variance for all pause times.

Figure 4 suggests that to generalize the differences between participants over trials, pause times within words required fewer trials (sentences) and items per trial (words per sentence) than pause times between words. Within-word pauses tended to be more independent of the trials and instead appeared to be related to participant behaviour. However, it is important to note that when all words were taken into account, the relative share of the error variance in comparison to the variances of the other facets was larger for pause times within words than between words. Moreover, the relative share of the error variance always made up a considerable part of the total variance regardless of the variable studied. This implies that when measuring pause times, a large part of the variance can be explained by the interaction of words and sentences combined with measurement errors. These large error variances therefore impact generalizability and, hence, the number of trials necessary to obtain measurements

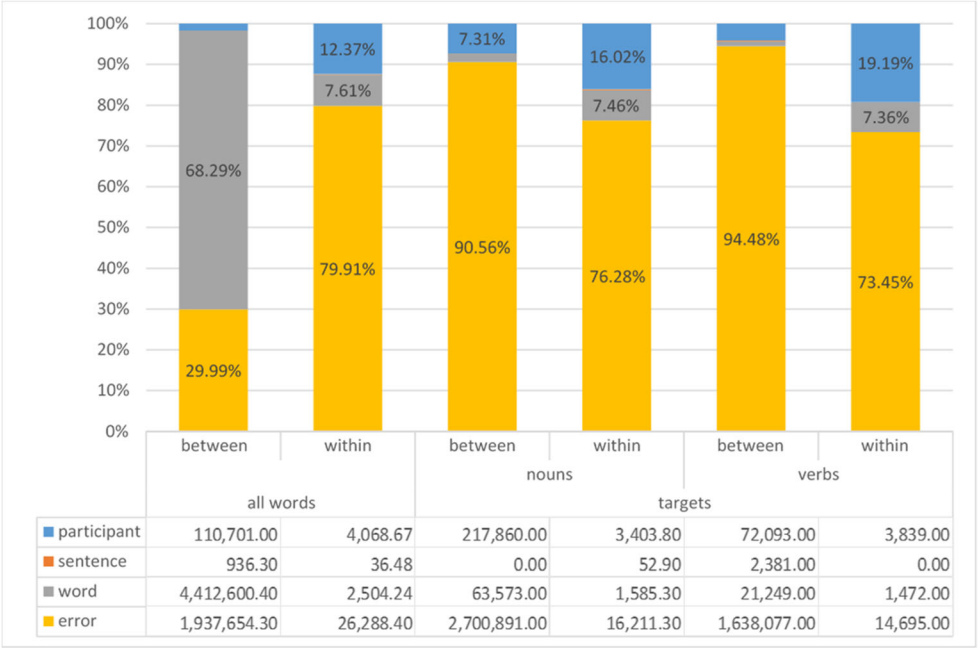


Fig. 4 Variance component estimations of pause times for all words and for target words for three sources of variance: participant, sentence and word. *Note*. The relative share of sentence in the total variance never exceeds 0.25% and is therefore not visible in the graph. Percentages below 5.00% are not shown

which can reliably inform statements about differences between participants. The online walkthrough (https://gifted-nightingale-d45b7e.netlify.app) also explains how the variances in Fig. 4 are estimated in R and how these variance estimates can be used to calculate the generalizability coefficients.

Generalizability study

To approximate the generalizability of the pause times, the variances (S^2) associated with the components (see Fig. 4), the number of fluently written sentences ($N_{sentence}$) and the number of words (N_{word}) were entered in Formula 3. Table 1 provides an overview of the generalizability scores for between- and within-word pause times for all words. The generalizability of both variables largely exceeded the .70 threshold. Consequently, the current number of sentences and the average number of words per sentence are more than sufficient to reliably draw conclusions on differences between participants.

The images and the verb infinitive that were displayed during each sentence production trial elicited target nouns and verbs. In the next step, the generalizability scores for these target words were separately estimated. For this, the variances (S^2) associated with the components (see Fig. 4) and the number of fluently written sentences ($N_{sentence}$) were re-entered in Formula 3. The mean number of words per sentence (N_{word}) was replaced by the mean number of targets.

Table 2 provides an overview of the generalizability scores for pause times between and within target words. The generalizability scores for pause times between nouns, within nouns and within verbs all exceeded the .70 threshold. However, the generalizability score of pause times between verbs did not reach the .70 threshold. This implies that, for pause times between verbs, the current number of sentences and the average number of words per sentence are not sufficient to reliably draw conclusions on differences between participants.

We will elaborate on these findings in the next section by exploring how much we can lower or need to increase the number of sentences and/or words per sentence to reach the .70 threshold.

Table 1 Generalizability scores for pause times between and within all words

Pause time	Sentences (*n*)	Words (*n*)	Generalizability
Between words	27	4.75	.83*
Within words	27	4.75	.95*

* ≥ .70.

Table 2 Generalizability scores for pause times between and within target words

Pause time	Sentences (*n*)	Targets (*n*)	Generalizability
Between nouns	27	1.75	.79*
Within nouns	27	1.75	.90*
Between verbs	27	1	.54
Within verbs	27	1	.87*

* ≥ .70.

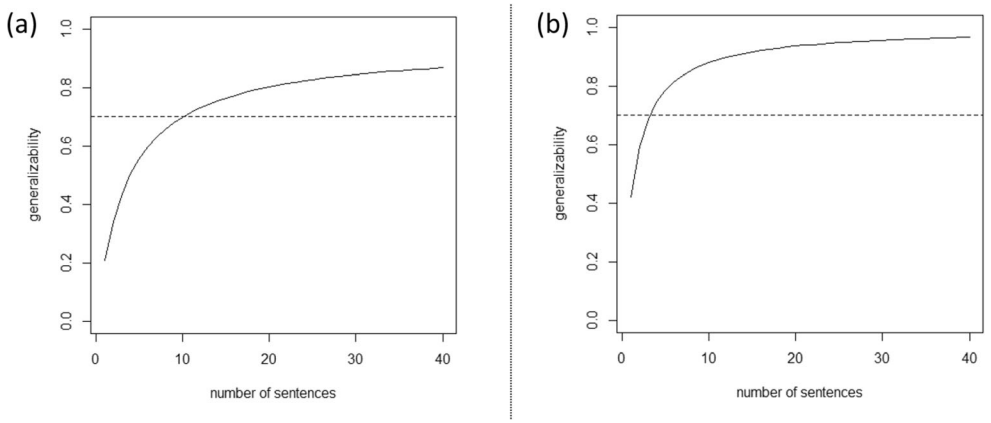


Fig. 5 Generalizability for pause times (a) between and (b) within words for all words, with an average of 4.75 words per sentence

Decision study

The approximation of the current task served as a starting point to extrapolate to new situations. For the variables of which the generalizability scores of the current task exceeded .70, it was estimated whether participants can be distinguished from one another with fewer sentences (trials)and/or fewer words (items) per sentence. For the pause time between verbs, the number of sentences and words that need to be added to the current task were estimated.

All words

For all words, the possibility of reducing the number of trials was explored by reducing the number of sentences in Formula 3. Results showed that pause times between words could be reliably estimated with 11 sentences ($E\rho^2$ = .72) if the mean number of words per sentence remained the same. As shown in Fig. 5, the exact cut-off point for the pause times within words was four sentences ($E\rho^2$ = .74).

Pause times were reanalysed by leaving out the ditransitive structures. With eight words per sentence, this was the longest of all structures. If participants only typed intransitive and monotransitive sentences, the mean sentence length would have dropped to 3.67 words. As Fig. 6 shows, the cut-off point of the pause time between words increased from 11 to 14 sentences ($E\rho^2$ = .71) when the number of words per sentence was reduced. However, the number of sentences for the pause times within words only increased with one: five sentences ($E\rho^2$ = .74) were necessary in order for $E\rho^2 \geq .70$.

Target words

For target words, Formula 3 was rewritten by lowering the number of sentences while keeping the mean number of target nouns fixed at 1.75. Results showed that reducing the number of trials in the sentence production task could not be done as extensively for nouns as for words in general. Pause times between nouns could only be reliably estimated with at least 17 sentences ($E\rho^2$ = .70) if the mean number of target nouns remained the same. For the pause time within nouns, the cut-off point was seven sentences ($E\rho^2$ = .72) (Fig. 7).

To explore a reduction of the mean number of target nouns per sentence, the ditransitive structures were left out. This means that the mean number of target nouns decreased from 1.75 to 1.33. Results indicated that 22 sentences ($E\rho^2$ = .70)

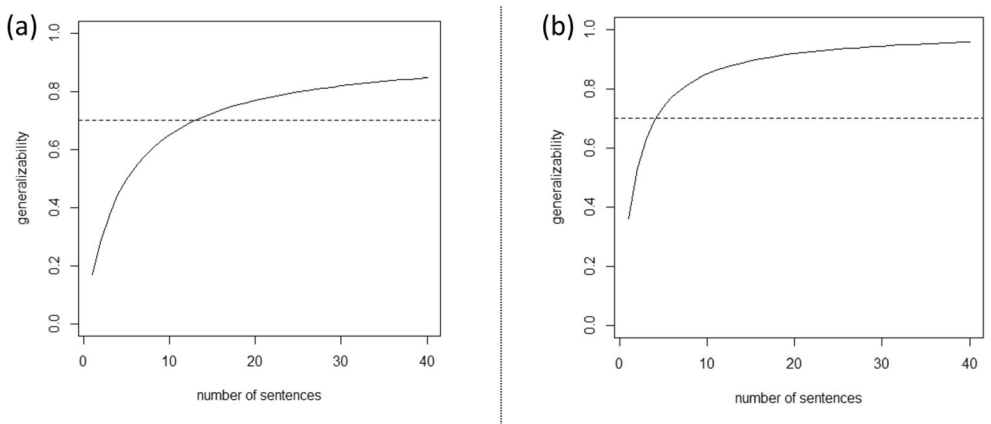


Fig. 6 Generalizability for pause times (a) between and (b) within words for all words, with an average of 3.67 words per sentence

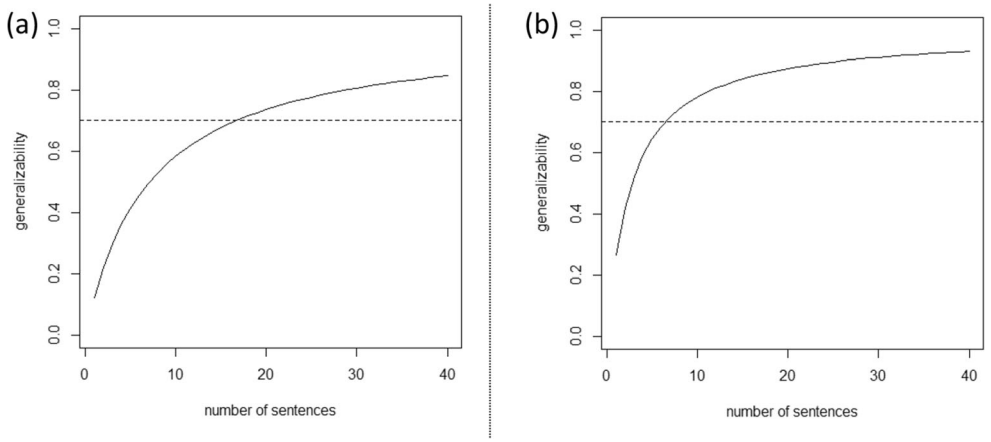


Fig. 7 Generalizability for pause times (a) between and (b) within nouns for target words, with an average of 1.75 nouns per sentence

instead of 17 were necessary to reliably estimate the pause time between nouns. For pause times within nouns, only nine sentences ($E\rho^2$ = .71) were needed (Fig. 8).

Similar to the approach for target nouns, the number of sentences in Formula 3 was changed as to examine the minimum number of sentences necessary to reliably draw conclusions on participants on the basis of pause times related to the verbs. The number of targets per sentence was not reduced because each sentence only contained one verb. Results indicated that 55 sentences ($E\rho^2$ = .70) were needed for the pause time between verbs, whereas the cut-off point was ten sentences ($E\rho^2$ = .72) for pause times within verbs (as shown in Fig. 9).

Discussion

The aim of the current paper was to demonstrate how generalizability theory can guide a researcher in determining (1) the minimum number of trials and (2) the minimum length of those trials needed in an experiment. For illustrative purposes, a sentence production task was used for which the minimum number of sentences and the minimum number of target words per sentence were estimated to reliably draw conclusions on differences between the writing processes of participants. Data on 3,180 fluently written sentences were collected. Pause times between and within words, nouns and verbs were the subject of analysis.

Results indicate that the ideal number of trials largely depends on the variables researchers want to draw conclusions from. For example, in the sentence production task, pause times within all words can be reliably estimated with four sentences, provided that the average number of words per sentence is set to 4.75. If the average number of words is reduced to 3.67, the cut-off point for pause times within words increases to five sentences. By contrast, for pause times between target verbs, 55 sentences are needed.

If multiple variables are used to generalize over trials, the chosen number of trials should allow for reliably estimating differences between participants on all of those variables. In this study, this would mean that a minimum number of 55 fluently typed sentences is necessary. This number only includes the trials in which no revisions are expected. Hence, to gather data on 55 fluently typed sentences, a total of 81 sentences would be needed if data loss remained 31.47% for

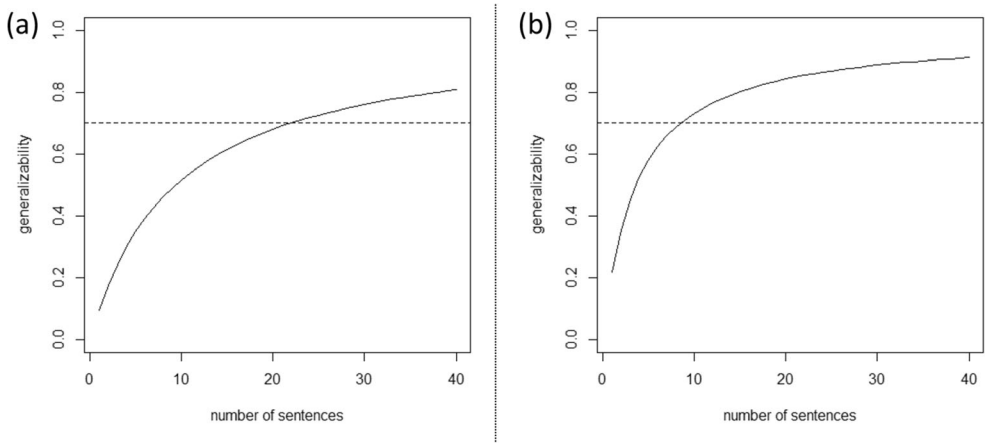


Fig. 8 Generalizability for pause times (a) between and (b) within nouns for target words, with an average of 1.33 nouns per sentence

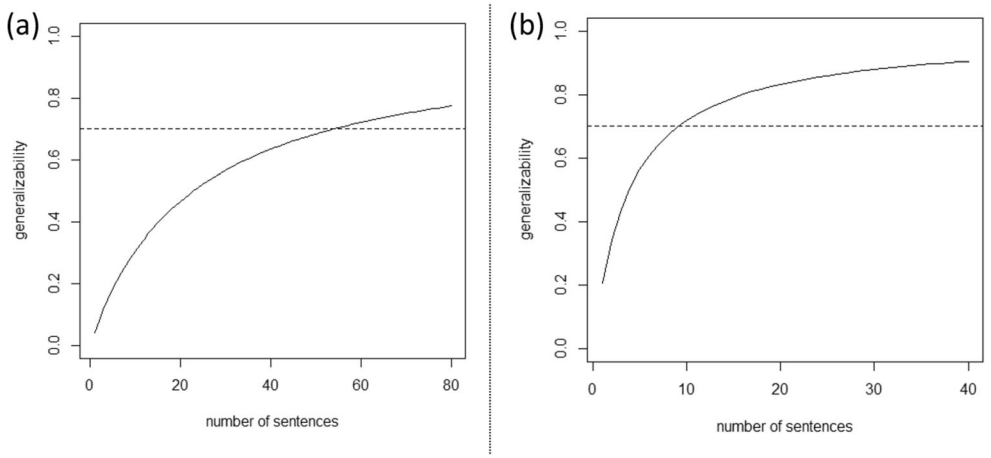


Fig. 9 Generalizability for pause times (a) between and (b) within verbs for target words

future data collection as it was in the current data collection (see data collection and analysis sections). However, other studies might perform the G and D studies on the complete data set, depending on the data that will be included in the analyses. In that case, compensation for data loss is not necessary.

If the minimum number of trials that is needed to generalize over trials on all variables increases participant burden too much, one could decide to only select those variables for which fewer trials are needed. In this study, for example, pause times between verbs require the highest number of sentences. If this variable is disregarded, the second highest number of sentences is sufficient to generalize over trials on the basis of the other remaining variables. Since it would be ideal to also reduce the number of items per trials, the minimum number of fluently written sentences would then be 22 (or a total of 33 when compensating for data loss). Of course, if the variable that requires the highest number of trials is crucial for the purpose of the experiment, it cannot be disregarded and the design of the experiment itself must be reconsidered.

It is also important to note that, in our example, reducing the number of items per trial (i.e., mean number of words per sentence) means that we are omitting one of the sentence structures (i.e., the ditransitive sentence structure, see Fig. 2) from our design. More specifically, if we want to lower the mean number of words per sentence from 4.75 to 3.67, we need to omit this ditransitive structure because it contains the largest number of words (i.e., eight words). However, if we still want to explore the writing processes related to the ditransitive structure in the follow-up studies, we should keep the current number of items per trial at 4.75 and adjust the number of trials accordingly.

A generalizability study also informs researchers about their variables' sensitivity. For example, in the sentence production task, pause times within words seem more sensitive for distinguishing participants than pause times between words. This can be derived from the fact that pause times within words require a much smaller number of trials to distinguish between participants. Identifying the most sensitive variables can be valuable for studies in which the aim is to distinguish groups based on those variables, as opposed to making statements about the variables themselves. This is similar to machine learning studies that identify those features that contribute most to assigning participants to a particular class (Toledo et al., 2014).

The added value of a G study depends on the quality of the data on which it is based. In order to be of value to a researcher, it is important to have sufficient observation units so that the variances are properly estimated and, hence, are reliable enough. Particularly when smaller data sets are used as the basis for a G study, it is valuable to highlight the uncertainty of estimates. This can be done by including confidence intervals making use of the Bayesian framework to estimate posterior distributions for the variances parameters (Lambert, 2018). However, in this article our aim was to only illustrate the concept and added value of a G study without adding the complexity of explaining Bayesian analyses and associated workflows. Future research could direct specific attention to integrating the concept of generalizability theory and Bayesian analyses.

It is important to note that the diversity of participants may influence the results of the G studies. For the current paper, we only used data of participants with no prior diagnosis of cognitive impairment. However, in the subsequent studies we aim to recruit both healthy participants and patients previously diagnosed with Alzheimer's disease. This brings us to a certain challenge that arises when performing a G study on data for vulnerable target groups. On the one hand, vulnerable groups are often heterogeneous, which makes it difficult to generalize the results of a G study from one sample to the next. To maximize this limited generalizability, researchers should, hence, use the largest possible data set for the G study. On the other hand, these groups are precisely the ones for

which it is rather difficult to find suitable participants and, as a result, it will not always be possible to base the G study on a large data set. Therefore, if researchers cannot use suitable data from the same target group as that of the subsequent studies, we recommend using data from a more homogeneous group (such as healthy participants with the same socio-demographic background instead of patients). To distinguish between participants that are more alike, it can be expected that more observations are needed than to distinguish between participants that differ greatly from each other. Hence, when data from a more homogeneous group are used, it is likely that the minimum number of trials is overestimated and that for subsequent studies even fewer observations are needed than the G study indicates. This does not exactly lead to the lowest number of trials, but it does ensure the usability of the data that are collected during subsequent studies.

The previous point highlights the difficulty in generalizing the results of G and D studies to experiments other than the ones they were based on. The task itself and the target group can, for instance, substantially influence the number of trials or items per trial that are needed (Bouwer et al., 2015; Schoonen, 2005). Therefore, researchers should repeat the analyses for each experiment to estimate how many tasks (and trials per task) are necessary. Similarly, van den Bergh et al. (2012) estimated the minimum number of writing tasks necessary to assess students' writing skills. However, differences in scoring procedure, language and student population influenced their results. For instance, the variance component estimation related to the writer facet was greater for first-year university students than for ninth-grade students. Hence, more written texts per student were required for the latter group to be able to generalize over trials and raters.

Finally, acceptable generalizability for this study was set at .70 higher. However, while there are guidelines for choosing a particular cut-off, there is no one-size-fits-all approach. A distinction is sometimes made between a generalizability score of .80 or higher, which is particularly suitable for studies with high-stakes assessments, and a generalizability score between .60 and .70, which is recommended for more formative assessments (Bloch & Norman, 2012). Hence, if the sentence production task presented here would, for example, have been put forward as a screening instrument to detect cognitive decline in a clinical setting rather than a measurement tool in a research setting, a higher cut-off of .80 or greater might have been better.

Conclusion

Our findings emphasize that, in comparison to more subjective methods of task reduction, G theory could function as an empirical framework to shorten task length and hence lower the burden on test subjects. This is especially useful in studies in which lowering the participant burden is of great importance, such as studies on dyslexia (Altemeier et al., 2008; Berninger et al., 2006; Wengelin & Strömqvist, 2000), aphasia (Behrns et al., 2010; Johansson-Malmeling, 2019; Johansson-Malmeling et al., 2021) and Alzheimer's disease (Afonso et al., 2019; Van Waes et al., 2017). However, results sometimes indicate that the number of trials should be increased instead of reduced. Therefore, G theory is a tool to estimate the minimum number of trials so that an experiment can be adjusted accordingly, whether this implies decreasing or increasing the number of trials.

Acknowledgements We would like to thank Johan Frid and Victoria Johansson for helping us with the development of a ScriptLog module that was perfectly tailored to our needs. Their insights in creating the module were indispensable. We would also like to thank all the elderly individuals who were willing to free up some of their valuable time to participate in our experiment.

Funding This work was supported by the Research Foundation – Flanders [Aspirant 2016, reference number: 1173617N, PeopleSoft ID: 3330].

References

Afonso, O., Álvarez, C. J., Martínez, C., & Cuetos, F. (2019). Writing difficulties in Alzheimer's disease and mild cognitive impairment. *Reading and Writing*, *32*, 217–233. https://doi.org/10.1007/s11145-017-9813-6

Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S., & McNamara, D. S. (2016). Investigating boredom and engagement during writing using multiple sources of information: the essay, the writer, and keystrokes. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 114–123. https://doi.org/10.1145/2883851.2883939

Altemeier, L. E., Abbott, R. D., & Berninger, V. W. (2008). Executive functions for reading and writing in typical literacy development and dyslexia. *Journal of Clinical and Experimental Neuropsychology*, *30*(5), 588–606. https://doi.org/10.1080/13803390701562818

Alves, R. A., Castro, S. L., & Olive, T. (2008). Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International Journal of Psychology*, *43*(6), 969–979. https://doi.org/10.1080/00207590701398951

American Psychological Association. (2002). Ethical Principles of Psychologists and Code of Conduct. *American Psychologist*, *57*(12), 1060–1073. https://doi.org/10.1037/0003-066X.57.12.1060

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Behrns, I., Ahlsén, E., & Wengelin, Å. (2010). Aphasia and text writing. *International Journal of Language & Communication Disorders*, *45*(2), 230–243. https://doi.org/10.3109/13682820902936425

Berninger, V. W., Abbott, R. D., Thomson, J., Wagner, R., Swanson, H. L., Wijsman, E. M., & Raskind, W. (2006). Modeling Phonological Core Deficits Within a Working Memory Architecture in Children and Adults With Developmental Dyslexia. *Scientific Studies of Reading*, *10*(2), 165–198. https://doi.org/10.1207/s1532799xssr1002_3

Bertram, R., Tønnessen, F. E., Strömqvist, S., Hyönä, J., & Niemi, P. (2015). Cascaded processing in written compound word production. *Frontiers in Human Neuroscience*, *9*, 207. https://doi.org/10.3389/fnhum.2015.00207

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68 *Medical Teacher*, *34*(11), 960–992. https://doi.org/10.3109/0142159X.2012.703791

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83–100. https://doi.org/10.1177/0265532214542994

Bouwer, R., Koster, M., & van den Bergh, H. (2018). Effects of a strategy-focused instructional program on the writing quality of upper elementary students in the Netherlands. *Journal of Educational Psychology*, *110*(1), 58–71. https://doi.org/10.1037/edu0000206

Bradburn, N. M. (1978). Respondent burden. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 35–40.

Brennan, R. L. (1992). Generalizability Theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brennan, R. L. (2001). *Generalizability Theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80–84. https://doi.org/10.1016/j.actpsy.2014.04.010

Cheong, C. M., Zhu, X., Li, G. Y., & Wen, H. (2019). Effects of intertextual processing on L2 integrated writing. *Journal of Second Language Writing*, *44*, 63–75. https://doi.org/10.1016/j.jslw.2019.03.004

Cohen, L., Manion, L., & Morrison, K. (2017). *Research Methods in Education* (8th ed.). Routledge.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley.

De Smedt, F., Van Keer, H., & Merchie, E. (2016). Student, teacher and class-level correlates of Flemish late elementary school children's writing performance. *Reading and Writing*, *29*, 833–868. https://doi.org/10.1007/s11145-015-9590-z

Fidalgo, R., Torrance, M., Rijlaarsdam, G., van den Bergh, H., & Lourdes Álvarez, M. (2015). Strategy-focused writing instruction: Just observing and reflecting on a model benefits 6th grade students. *Contemporary Educational Psychology*, *41*, 37–50. https://doi.org/10.1016/j.cedpsych.2014.11.004

Ford, M., & Holmes, V. M. (1978). Planning units and syntax in sentence production. *Cognition*, *6*(1), 35–53. https://doi.org/10.1016/0010-0277(78)90008-2

Frid, J., Johansson, V., Johansson, R., & Wengelin, Å. (2014). *Developing a keystroke logging program into a writing experiment environment*. Abstract from Writing Across Borders.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, *26*(4), 507–531. https://doi.org/10.1177/0265532209340188

Graham, S., Daley, S. G., Aitken, A. A., Harris, K. R., & Robinson, K. H. (2018). Do writing motivational beliefs predict middle school students' writing performance? *Journal of Research in Reading*, *41*(4), 642–656. https://doi.org/10.1111/1467-9817.12245

Graham, S., Hebert, M., Paige Sandbank, M., & Harris, K. R. (2016). Assessing the Writing Achievement of Young Struggling Writers. *Learning Disability Quarterly*, *39*(2), 72–82. https://doi.org/10.1177/0731948714555019

Johansson-Malmeling, C. (2019). *Changes in writing processes caused by post-stroke aphasia or low-grade glioma* [University of Gothenburg. Sahlgrenska Academy]. http://hdl.handle.net/2077/61828

Johansson-Malmeling, C., Hartelius, L., Wengelin, Å., & Henriksson, I. (2021). Written text production and its relationship to writing processes and spelling ability in persons with post-stroke aphasia. *Aphasiology*, *35*(5), 615–632. https://doi.org/10.1080/02687038.2020.1712585

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1–26. https://doi.org/10.17239/jowr-2008.01.01.1

Kellogg, R. T. (2004). Working Memory Components in Written Sentence Generation. *The American Journal of Psychology*, *117*(3), 341–361. https://doi.org/10.2307/4149005

Knospe, Y., Sullivan, K. P. H., Malmqvist, A., & Valfridsson, I. (2019). Observing Writing and Website Browsing: Swedish Students Write L3 German. In E. Lindgren & K. P. H. Sullivan (Eds.), *Observing Writing: Insights from Keystroke Logging and Handwriting* (pp. 258–284). Brill. https://doi.org/10.1163/9789004392526_013

Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. SAGE Publications Ltd.

Leijten, M., De Maeyer, S., & Van Waes, L. (2011). Coordinating sentence composition with error correction: A multilevel analysis. *Journal of Writing Research*, *2*(3), 331–363. https://doi.org/10.17239/jowr-2011.02.03.3

Leijten, M., Van Horenbeeck, E., & Van Waes, L. (2019). Analysing Keystroke Logging Data from a Linguistic Perspective. In E. Lindgren & K. P. H. Sullivan (Eds.), *Observing Writing: Insights from Keystroke Logging and Handwriting* (pp. 71–95). BRILL. https://doi.org/10.1163/9789004392526_005

Leijten, M., & Van Waes, L. (2013). Keystroke Logging in Writing Research. *Written Communication*, *30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Lingler, J. H., Schmidt, K. L., Gentry, A. L., Hu, L., & Terhorst, L. A. (2014). A New Measure of Research Participant Burden. *Journal of Empirical Research on Human Research Ethics*, *9*(4), 46–49. https://doi.org/10.1177/1556264614545037

Mateos, M., & Solé, I. (2009). Synthesising information from various texts: A study of procedures and products at different educational levels. *European Journal of Psychology of Education*, *24*(4), 435–451. https://doi.org/10.1007/BF03178760

Medimorec, S., & Risko, E. F. (2017). Pauses in written composition: on the importance of where writers pause. *Reading and Writing*, *30*, 1267–1285. https://doi.org/10.1007/s11145-017-9723-7

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods*, *38*(3), 542–547. https://doi.org/10.3758/BF03192810

Negro, I., Chanquoy, L., Fayol, M., & Louis-Sidney, M. (2005). Subject-Verb Agreement in Children and Adults: Serial or Hierarchical Processing? *Journal of Psycholinguistic Research*, *34*(3), 233–258. https://doi.org/10.1007/s10936-005-3639-0

Nottbusch, G. (2010). Grammatical planning, execution, and control in written sentence production. *Reading and Writing*, *23*, 777–801. https://doi.org/10.1007/s11145-009-9188-4

Nottbusch, G., Weingarten, R., & Sahel, S. (2007). From Written Word to Written Sentence Production. In M. Torrance, L. van Waes, & D. Galbraith (Eds.), *Writing and Cognition: Research and Applications* (Vol. 20, pp. 31–53). Brill. https://doi.org/10.1163/9781849508223_004

Paesen, L. & Meulemans, C. (2020). *Open Linguistic Picture Database* [Data set], Zenodo, https://doi.org/10.5281/zenodo.3738213

Power, M. J. (1985). Sentence Production and Working Memory. *The Quarterly Journal of Experimental Psychology Section A*, *37*(3), 367–385. https://doi.org/10.1080/14640748508400940

Purcell, J. J., Turkeltaub, P. E., Eden, G. F., & Rapp, B. (2011). Examining the Central and Peripheral Processes of Written Word Production Through Meta-Analysis. *Frontiers in Psychology*, *2*, 239. https://doi.org/10.3389/fpsyg.2011.00239

Quinlan, T., Loncke, M., Leijten, M., & Van Waes, L. (2012). Coordinating the Cognitive Processes of Writing: The Role of the Monitor. *Written Communication*, *29*(3), 345–368. https://doi.org/10.1177/0741088312451112

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Scaltritti, M., Arfé, B., Torrance, M., & Peressotti, F. (2016). Typing pictures: Linguistic processing cascades into finger movements. *Cognition*, *156*, 16–29. https://doi.org/10.1016/j.cognition.2016.07.006

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing*, *22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. SAGE Publications Inc.

Shavelson, R. J., & Webb, N. M. (2006). Generalizability Theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of Complementary Methods in Education Research* (pp. 309–322). Routledge.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922–932. https://doi.org/10.1037/0003-066X.44.6.922

Sullivan, K., & Lindgren, E. (Eds.). (2006). *Computer Keystroke Logging and Writing*. BRILL. https://doi.org/10.1163/9780080460932

Toledo, C. M., Cunha, A., Scarton, C., & Aluísio, S. (2014). Automatic classification of written descriptions by healthy adults: An overview of the application of natural language processing and machine learning techniques to clinical discourse analysis. *Dementia & Neuropsychologia*, *8*(3), 227–235. https://doi.org/10.1590/S1980-57642014DN83000006

Troia, G. A., Harbaugh, A. G., Shankland, R. K., Wolbers, K. A., & Lawrence, A. M. (2013). Relationships between writing motivation, writing activity, and writing performance: effects of grade, sex, and ability. *Reading and Writing*, *26*, 17–44. https://doi.org/10.1007/s11145-012-9379-2

Ulrich, C. M., Wallen, G. R., Feister, A., & Grady, C. (2005). Respondent Burden in Clinical Research: When Are We Asking Too Much of Subjects? *IRB: Ethics and Human Research*, *27*(4), 17–20. https://doi.org/10.2307/3563957

van den Bergh, H., De Maeyer, S., van Weijen, D., & Tillema, M. (2012). Generalizability of Text Quality Scores. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring Writing: Recent Insights into Theory, Methodology and Practice* (pp. 23–32). BRILL. https://doi.org/10.1163/9789004248489_003

Van Waes, L., Leijten, M., Mariën, P., & Engelborghs, S. (2017). Typing competencies in Alzheimer's disease: An exploration of copy tasks. *Computers in Human Behavior*, *73*, 311–319. https://doi.org/10.1016/j.chb.2017.03.050

van Weijen, D., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2009). L1 use during L2 writing: An empirical study of a complex phenomenon. *Journal of Second Language Writing*, *18*(4), 235–250. https://doi.org/10.1016/j.jslw.2009.06.003

Weingarten, R., Nottbusch, G., & Will, U. (2004). Morphemes, syllables and graphemes in written word production. In T. Pechmann & C. Habel (Eds.), *Multidisciplinary Approaches to Language Production* (pp. 529–572). De Gruyter Mouton. https://doi.org/10.1515/9783110894028.529

Wengelin, Å. (2006). Examining Pauses in Writing: Theory, Methods and Empirical Data. In K. Sullivan & E. Lindgren (Eds.), *Computer Keystroke Logging and Writing: Methods and Applications* (pp. 107–130). : Elsevier.

Wengelin, Å., & Strömqvist, S. (2000). Discourse level writing in dyslexics – methods, results, and implications for diagnosis. *Logopedics Phoniatrics Vocology*, *25*(1), 22–28. https://doi.org/10.1080/140154300750045876

Wengelin, Å., Torrance, M., Holmqvist, K., Simpson, S., Galbraith, D., Johansson, V., & Johansson, R. (2009). Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior Research Methods*, *41*(2), 337–351. https://doi.org/10.3758/BRM.41.2.337