Research Methods in Applied Linguistics 2 (2023) 100079

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

GAM-based individual difference measures for L2 ERP studies

Nienke Meulman [a,*], Simone A. Sprenger [a], Monika S. Schmid [b], Martijn Wieling [a]

[a] Center for Language and Cognition, University of Groningen, Groningen, the Netherlands
[b] Department of Language and Linguistic Science, University of York, York, UK

ARTICLE INFO

Keywords:
Bilingual development
Grammatical processing
Statistical method
Individual differences
Event-related potentials

ABSTRACT

ERPs (Event-Related Potentials) have become a widely-used measure to study second language (L2) processing. To study individual differences, traditionally a component outcome measure is calculated by averaging the amplitude of a participant's brain response in a pre-specified time window of the ERP waveform in different conditions (e.g., the 'Response Magnitude Index'; Tanner, Mclaughlin, Herschensohn & Osterhout, 2013). This approach suffers from the problem that the definition of such time windows is rather arbitrary, and that the result is sensitive to outliers as well as participant variation in latency. The latter is particularly problematic for studies on L2 processing. Furthermore, the size of the ERP response (i.e., amplitude difference) of an L2 speaker may not be the best indicator of near-native proficiency, as native speakers also show a great deal of variability in this respect, with the 'robustness' of an L2 speaker's ERP response (i.e., how consistently they show an amplitude difference) potentially being a more useful indicator. In this paper we introduce a novel method for the extraction of a set of individual difference measures from ERP waveforms. Our method is based on participants' complete waveforms for a given time series, modelled using generalized additive modelling (GAM; Wood, 2017). From our modelled waveform, we extract a set of measures which are based on amplitude, area and peak effects. We illustrate the benefits of our method compared to the traditional Response Magnitude Index with data on the processing of grammatical gender violations in 66 Slavic L2 speakers of German and 29 German native speakers. One of our measures in particular appears to outperform the others in characterizing differences between native speakers and L2 speakers, and captures proficiency differences between L2 speakers: the 'Normalized Modelled Peak'. This measure reflects the height of the (modelled) peak, normalized against the uncertainty of the modelled signal, here in the P600 search window. This measure may be seen as a measure of peak robustness, that is, how reliable the individual is able to show a P600 effect, largely independently of where in the P600 window this occurs. We discuss implications of our results and offer suggestions for future studies on L2 processing. The code to implement these analyses is available for other researchers.

* Corresponding author at: Center for Language and Cognition, University of Groningen, Oude Kijk in 't Jatstraat 26, P.O. Box 716, 9700 AS Groningen, the Netherlands.
E-mail address: n.meulman@rug.nl (N. Meulman).

https://doi.org/10.1016/j.rmal.2023.100079
Received 29 January 2023; Received in revised form 31 August 2023; Accepted 31 August 2023
Available online 15 September 2023

1. Introduction

Second language (L2) learning is subject to substantial individual variation. However, traditional Event-Related Potential (ERP) studies on L2 learning are often restricted to group comparisons of the grand mean waveforms of L2 speakers versus native speakers. This approach stems from methodological considerations, since ERPs have to be averaged over both trials and individuals in order to achieve adequate signal-to-noise ratio. In recent years, however, there has been increasing interest in individual differences in neurocognitive aspects of L2 processing. While this change of direction in the field in itself is positive, it brings with it some methodological challenges and concerns. Here, we present a method to extract a reliable individual measure from the ERP signal that is better able to deal with individual differences than the traditional measure.

1.1. ERPs in L2 processing and individual differences

ERPs are derived from the electroencephalogram (EEG) and offer a powerful tool to investigate online language processing (see Luck, 2014, for an introduction to ERPs). Their fine-grained temporal resolution allows the investigation of real-time language processing, and thus insights into similarities and differences between languages acquired from birth in (monolingual) native speakers and languages acquired later in life (L2s) (see Steinhauer, 2014, and Morgan-Short, 2014, for overviews of ERP research on L2 acquisition).

The current paper introduces a methodology for the analysis of individual differences in language processing as investigated through the ERP method in order to gain further insight into similarity or difference between native and advanced L2 processing. For the purpose of this study, we focus on a particular ERP component, namely the P600, which is a large positive-going ERP wave with a maximum around 600 ms post stimulus onset commonly and robustly observed in response to grammatical violations,[1] but the procedure we propose is equally suitable for other (language-related) ERP components, such as the N400 or the LAN (see Payne, Ng, Shantz & Federmeier, 2020, for a review of the ERP components that are central in the study of multilingual language processing). Current theories of the P600 interpret its functional significance as reflecting a process of revision, reanalysis or integration difficulty (e.g., Brouwer, Crocker, Venhuizen & Hoeks, 2017; Kaan & Swaab, 2003; Kos, Vosse, Van Den Brink & Hagoort, 2010). A modulation of this component in L2 speakers, compared to native speakers, may inform us about differences in linguistic processing between the two groups. Our focus here is on how to extract a reliable measure representing an individual's brain response to grammatical violations, which will allow better insight into individual differences between L2 speakers.

Most ERP research in the field of L2 grammar processing has been conducted using group comparisons (e.g., L2 speakers vs. native speakers, low proficiency vs. high proficiency L2 speakers, or early vs. late onset L2 speakers), in which participant variability is treated as a source of noise (e.g., in the error term in an ANOVA model). ERPs are used to investigate to what extent a particular group of L2 speakers can acquire new grammatical features, i.e., whether they are able to show the same neurocognitive response to manipulations of these features as native speakers do, and under which circumstances they show quantitative and/or qualitative deviations from the native norm (e.g., Alemán Bañón, Fiorentino, & Gabriele, 2018; Carrasco-Ortíz, Herrera, Jackson-Maldonado, Ramírez, Pereyra & Wicha, 2017; Foucart & Frenck-Mestre, 2012; Meulman, Stowe, Sprenger, Bresser, & Schmid, 2014; Morgan--Short, Sanz, Steinhauer, & Ullman, 2010). The finding of a less pronounced/delayed/absent/different ERP effect is then taken to reflect weaker/slower/different neurocognitive mechanisms or neural processing, for example because the syntactic structure under investigation is not fully acquired (yet), or stored differently in the language system of the L2 speaker compared to that of a native speaker.

More recently, other studies have opted to make individual differences in this respect the subject of their investigation (e.g., Alemán Bañón, Miller, & Rothman, 2017; Bice & Kroll, 2021; Bond, Gabriele, Fiorentino, & Alemán Bañón, 2011; Grey, 2022; Grey, Tanner & van Hell, 2017; Kim, Oines & Miyake, 2018; Tanner, Inoue, & Osterhout, 2014; Tanner, Mclaughlin, Herschensohn, & Osterhout, 2013; Tanner, Goldshtein & Weissman, 2018; Tanner & van Hell, 2014; Tanner, 2019). These studies demonstrate the robust impact of a broad range of individual factors on both semantic and syntactic processing, as measured by ERPs. Individual variability exists in both monolingual and bilingual populations (Beatty-Martínez, Bruni, Bajo & Dussias, 2021; Tanner, Goldshtein & Weissman, 2018), with ERP responses modulated by general factors such as working memory span and handedness, but it is compounded in bilingual populations by the impact of L2-specific factors such as age of acquisition and proficiency. As a result, ERP outcomes presented in grand mean average waveforms and the accompanying omnibus statistical analyses of mean amplitudes may present one picture, but closer inspection of between-participant variation reveals another. For example, group grand averages have sometimes found biphasic patterns, suggesting the presence of two separate ERP components, but subsequent analysis reveals that in fact subsets of individuals exhibit only one of the components (see Tanner, Goldshtein & Weissman, 2018 for discussion). This is particularly relevant for investigations of L2 speakers at different stages of proficiency, where the biphasic pattern comprising both an N400 and a P600 may be due to the fact that the lower proficiency L2 speakers tend to exhibit the former in response to grammatical variations, while L2 speakers at higher proficiency levels exhibit a P600 (e.g., Osterhout, 1997; Tanner et al., 2014). Crucially, averaging over populations comprising both types of responses may cancel out characteristic patterns (Tanner, Goldsthein & Weissman, 2018). Other recent studies find similar contrasts between grand-average waveforms and results showing the full range of brain responses within particular L2 groups (Bice & Kroll, 2021; Grey, 2022).

These insights demonstrate that grand mean analyses can obscure some systematic variation between individuals in ERP responses. At present, we do not have a clear idea why some of this variation exists, nor how to incorporate these findings in current models of

[1] It is not the purpose of this paper to contribute to the discussion on the precise nature of the linguistic processes indexed by the P600, as recently referenced, for example, in Leckey & Federmeier (2019) or Sassenhagen & Fiebach (2019).

language processing. This research nevertheless emphasizes that it is important to take individual variation into account, particularly in (typically more heterogeneous) L2 populations, and that the challenge for future research is to develop more powerful ways to capture variation in all of the dimensions that ERPs have to offer in order to help us gain a better understanding of the underlying mechanisms.

In the current paper, we take a step in this direction. Analyses based on mean amplitude certainly have their merits (increased signal-to-noise ratio) and have contributed many interesting insights into real time L2 processing. They may, however, fail to recognize important between-participant differences, and can even lead to wrong conclusions. Therefore, it is important that more research should focus on individual variation in the ERP signal of L2 speakers. The goal of the current paper is to provide a new set of tools to study these individual differences.

1.2. Problems with the traditional Response Magnitude Index

Some of the studies of individual differences in L2 ERPs listed above used a measure of average response/effect magnitude (the Response Magnitude Index,[2] RMI, in Tanner, Inoue & Osterhout, 2014), calculated for each individual. This is defined as the mean activity difference between two conditions (e.g., ungrammatical minus grammatical condition in a morphosyntactic violation paradigm), averaged over all trials in a somewhat arbitrarily selected time window (usually 300–500 ms for the N400/LAN, and 500–800 to 1000 ms for the P600), for a particular selection of electrodes. This measure has been very valuable in allowing researchers to discover important patterns of individual variation (see Pélissier, 2020, for an overview). However, there are also a number of drawbacks associated with this approach.

The first two issues of the Response Magnitude Index concern the practice of *averaging over the items* and *averaging across a time window*. When averaging over items, outliers may have a substantial effect on the calculated average. This is particularly problematic in ERP data, which has a relatively low signal-to-noise ratio. Calculating ERP magnitude as the average amplitude of the ERP waveform in a certain time window has drawbacks (see Luck, 2014 and Luck & Gaspelin, 2017 for discussion, and see Meulman, Wieling, Sprenger, Stowe, & Schmid, 2015 for an approach to avoid these problems). In particular, there is no independent objective way to define the measurement window. Instead, the recommended standard practice is to select a time window based on previous reports and visual inspection of the data. As a consequence, participant variability with respect to latency is ignored, which may lead to a loss of power and may cause some L2 speakers' capacity to be underestimated.

This poses a particular problem for research on L2 acquisition and questions of the attainability of native-like grammatical processing. It has often been pointed out that the added cognitive load imposed by the management of two linguistic systems in bilinguals may lead to differences in accuracy rates (Hopp, 2010) and linguistic processing (Green, 2011). In particular, delayed responses (Bialystok, 2009) and slower processing routines (Kotz, 2009) have been found. This implies that even if bilinguals do show a similar effect in the ERP signal as the monolinguals in response to a violation in the input, it may occur at a different latency – and the fact that similar effects have often been observed in the native language of bilinguals (e.g., Schmid & Köpke, 2017) implies that they do not necessarily reflect differences between native speakers and L2 speakers but rather between monolinguals and bilinguals. This problem is further increased as the delay may be modulated by individual factors, such as proficiency level. Such differences are again likely to be greater in a bilingual than a monolingual population. Since averaging the signal across the same time-window for both populations may thus make the effect for L2 speakers appear smaller or even lead to its disappearance altogether, latency and amplitude effects should be teased apart.

Traditionally, latency effects have only been investigated in grand average waveforms of groups of L2 speakers (e.g., Kotz, Holcomb, & Osterhout, 2008; Rossi, Gugler, Friederici, & Hahne, 2006; Sabourin & Stowe, 2008). These studies do find delayed P600 effects in (lower proficiency) L2 speakers compared to native speakers, and attribute this to more uncertainty and/or processing problems for L2 speakers. To our knowledge, there have thus far not been any ERP studies that have examined latency effects of ERP components in L2 speakers at the individual level.

A final concern about the Response Magnitude Index as a gradient indicator of the native-likeness of responses to ungrammaticalities is more fundamental in nature. By definition, in this measure it is the magnitude (i.e., amplitude difference) of the ERP response of an L2 speaker that is taken as an indicator of nativelikeness. However, the actual size of an individual's ERP component is dependent on many factors, and varies considerably even in monolingual native speakers (who have full mastery of the language). The size of the P600, for example, has been shown to be influenced by factors such as cognitive control (Beatty-Martínez et al. 2021) or differences in linguistic processing that are yet to be further determined (Tanner, 2019). Of particular interest here is the finding that language experience plays a role in determining the size of the P600 (Pakulak & Neville, 2010). Since monolingual and bilingual populations by definition have different levels of exposure (as the bilinguals have to divide the available time between their languages), differences in amplitude are problematic in establishing whether L2 speakers are or are not 'nativelike', as a smaller average effect in an L2 population may simply be the result of these speakers not having been exposed to the structure as often as the native speakers. Arguably, therefore the *size* of an ERP response in an L2 speaker may not be the best indicator of nativelikeness, and a more appropriate measure may be its robustness (i.e., how reliably they are able to show the response we would generally find in a

[2] In addition to the RMI, Tanner et al. also calculated the Response Dominance Index (RDI), which is a metric of how N400- or P600-dominant a participant's brain response is. Although this is a useful measure, it is less relevant for the current paper, which focusses on amplitudes above the baseline when studying components with a positive polarity (e.g., P600), and amplitudes below the baseline for negative polarity components (e.g., N400).

native speaker).

To overcome the issues with the Response Magnitude Index, as well as to test our hypothesis that the robustness of an L2 speaker's ERP response provides a good measure of nativelikeness, we propose an approach that extracts a set of ERP component measures based on participants' complete waveforms for a given time series using generalized additive modelling (Wood, 2017).

2. Generalized additive modelling-based individual difference measures

2.1. Estimating ERPs through generalized additive modelling

As a first step, our approach entails using a generalized additive model (GAM; Wood, 2017) capable of estimating ERPs through a non-linear regression approach. The activity present in the ERP waveform recorded for a single trial (within a participant) will include stimulus-evoked potentials, as well as background activity that is not related to the stimulus. The standard ERP approach averages these observed voltages for multiple trials directly, and the resulting averages are used in a subsequent statistical analysis. Often, such analyses also require the dichotomozation of inherently interval variables, such as frequency. GAMs provide a nonparametric regression smooth fit of an individual's ERP response over time, while appropriately dealing with item-based variability through a mixed-effects regression approach (see Wood, 2017, for a complete overview of the theory and practical applications, and Meulman et al., 2015 and Abugaber et al., 2023, for examples of applications for L2 research).

The advantage of using a statistical approach to obtain the (smooth) ERP signal, rather than using an averaging procedure, is that the resulting modelled waveform is much less susceptible to outliers. GAMs are able to appropriately deal with the complex shape of the ERP signal, since they identify the non-linear pattern automatically without overfitting (see also the discussion in Wieling, 2018). Importantly, in finding the optimal fit, GAMs do not minimize the difference between the model fit and the actual values (i.e., the error), but rather minimize a combination of this error and a smoothness penalty. The consequence of this is that non-linear (i.e., less smooth) patterns are only obtained when adequately supported by the data (assessed via cross-validation). As generalized additive modelling is a regression-based approach, it is well suited to assess the influence of morphosyntactic manipulations on the ERP signal as it unfolds over time.

Since the goal of this paper is to identify participant-specific measures, we fit participant-specific GAMs on the basis of the time-locked EEG signal (i.e., per sentence and per participant) across the full time range per trial (of e.g., -500 to 1400 ms before/after target onset). A simple model specification (in R code) may look as follows:

$$\mu\mathrm{V} \sim \mathrm{s}(\mathrm{Time}) + \mathrm{s}(\mathrm{Time}, \mathrm{by} = \mathrm{IsUngrammatical}) + \mathrm{s}\left(\mathrm{Time}, \mathrm{Item}, \mathrm{bs} = \text{"fs"}, \mathrm{m} = 1\right)$$

This specification indicates that our dependent variable is the ERP amplitude (in microvolts), which we are modelling by allowing a non-linear effect of time (i.e. s(Time)), and we obtain the pattern for the difference waveform directly by including a binary difference smooth (ungrammatical: 1 vs. grammatical: 0). The final item in the model specification represents the random effect of item via a factor smooth (bs="fs"), which models the individual (non-linear) variability in the ERP signal across items. To see why the second term indeed models the (potential) non-linear difference between the grammatical and ungrammatical term, it is important to note that when the by-variable "IsUngrammatical" equals 0 (in the grammatical case), the second smooth is reduced to 0 (i.e., removed). In the other case, this smooth will model a non-linear pattern. Consequently, for the grammatical case, the fixed-effect pattern only consists of the first smooth over time, s(Time), whereas for the ungrammatical case the resulting pattern is modelled by the first and the second smooth over time added together. This means that the second smooth must model the difference between the grammatical and the ungrammatical condition (see also Wieling, 2018). The modelled difference waveform is used as a basis from which the different participant-specific measures are extracted.

Analysing the ERP signal through GAMs bears some similarity to other single-trial methods, particularly the rERP framework (Smith & Kutas, 2015a,b). The rERP approach uses regression modelling "to estimate the value of the ERP at one single electrode and latency" (Smith & Kutas, 2015a, p. 161). The result of this approach is an estimated ERP waveform, with the advantage that it also allows for (component) overlap correction, and the waveform can be analysed just as ERPs are (e.g., using averaging across a certain time window). While our approach does not allow for overlap correction, the shape of the GAM smooth does allow us to investigate the presence (and robustness) of a peak in this modelled (difference) ERP signal (see Section 2.2.3 for further discussion). In contrast to the rERP approach, the GAM approach does not assume independence of subsequent data points, and offers a simpler procedure to assess significance. In principle, our approach of extracting the set of individual difference measures described below would also be applicable using the rERP signal instead of the GAM difference waveform as input, as long as the standard error on the basis of the underlying trials is known.

2.2. Extracting the individual difference measures

After fitting a GAM for each participant, the second step of our method involves extracting a set of individual measures from the GAM smooth, which are based on amplitude, area and peak effects. In addition to the traditional Response Magnitude Index, we have created two other measures that intend to quantify the size of the response (i.e., the amount of neural activity), but which are less susceptible to influence by outliers and less dependent on a specific time window, and therefore likely offer a more reliable representation. In addition, we propose a measure that captures whether the response is robust: a consistent response with little variability across trials is more robust than one with substantial variability across trials. Finally, we have created two measures that can be used to

investigate latency effects in the timing of the ERP response.

The extracted measures are summarized in Table 1 and illustrated in Figs. 1 and 2, and will be further explained in the next subsections. For completeness, we also include the traditional Response Magnitude Index (1), which is based on the mean of the original data, in addition to our new measures, Modelled Area, Height Modelled Peak, Normalized Modelled Peak, Modelled Area Median Latency and Modelled Peak Latency (2–6), which are based on the GAM fitted smooth of the difference waveform.

While these measures are not intended to be able to quantify variation across all dimensions in which variation in ERP waveforms may present itself, they provide a (useful) starting point, allowing researchers to choose whichever best suits their research question and particular dataset. In the empirical example that we present below, we demonstrate the benefits and drawbacks of each of the extracted measures.

2.2.1. Traditional Response Magnitude Index

To evaluate the effectiveness of our GAM-based individual difference measures, we include the traditional Response Magnitude Index (*measure 1* in Table 1) of quantifying the average ERP response magnitude. As indicated, this measure has been used in previous research (e.g., Alemán Bañón et al., 2017; Tanner et al., 2014) and is defined as the mean activity difference between two conditions (e.g., ungrammatical minus grammatical) averaged over all trials, in a selected time window (e.g., 500–1000 ms), for a particular (selection of) electrode(s). The unit of measurement is amplitude in microvolt (μV), and the measure is intended to quantify the size of the response (i.e., the amount of neural activity) for the particular component under investigation. In the example presented in Fig. 1, the Response Magnitude Index would yield a value of around 3.6 μV.[3]

2.2.2. Modelled area measures

The Modelled Area (*measure 2* in Table 1) is based on 'signed area amplitude' as defined by Luck (2014), which has been put forward as an alternative to determining the mean amplitude over a time window. Specifically, similar to the claims that we have made in Section 1.2 of this paper, Luck states that if a component occurs later in one group than in another (which is often the case for L2 speakers compared to native speakers), it may be problematic to compare mean amplitude over the same time window in both groups. Luck suggests instead to use the signed area measure: the geometric area under the curve (with 'signed' meaning positive or negative, depending on whether one is interested in regions above or below the baseline), measured in units of μV * ms. This measure is not sensitive to differences in latency. The advantage of using this measure is that one can use a fairly wide measurement window, without any cancellation from previous or subsequent waves in the opposite direction (due to the 'signed' part of the measure). It also eliminates any bias that we might introduce by selecting a narrow time window on the basis of the observed time course of the effect.

In this paper, we differentiate between the terms 'time window' and 'search window', to mark this distinction between a time window in which all measurement points of the waveform are included in the calculation of the measure (such as in the Response Magnitude Index), versus a search window in which only measurement points of the waveform that meet certain criteria are taken along in the calculation of the measure (such as our Modelled Area measures, and also the Modelled Peak measures which we will discuss next).

Note that we calculate the Modelled Area based on the GAM fitted smooth, rather than on the original data. This eliminates some of the disadvantages of the signed area measure Luck (2014) mentions. Specifically, he warns that noisy waveforms will tend to have larger area values than clean waveforms, and consequently one cannot compare area values for groups or conditions with different noise levels. Using GAMs in combination with determining a difference smooth should largely alleviate this problem, as it will use the underlying smooth trajectory rather than the noisy signal to determine the area.

The Modelled Area measure was furthermore used to extract a measure of latency (*measure 5* in Table 1), by using the fractional area technique (Hansen & Hillyard, 1980). This technique defines the latency of the component as the first time point at which a certain percentage (typically, 50%) of the total area of the component has been reached (i.e., the median). Fractional area latency, traditionally computed from an area of the average waveform, tends to provide the most accurate method for measuring changes in latency of a range of ERP components (Kiesel, Miller, Jolicœur & Brisson, 2008). Here, we extract the 'Modelled Area Median Latency' by using the 50% area latency of the signed geometric area under the GAM difference smooth.

2.2.3. Modelled Peak measures

When researchers first started to use ERPs to investigate neurocognitive processing (i.e., roughly from the 1930s to the 1970s), they had to resort to measuring the size of the ERP with a ruler (Donchin & Heffly, 1978), and peaks were simply the easiest characteristic of the waveform to measure. Therefore, peaks were the standard measure at the time. Later, when computational techniques improved, people were able to calculate more sophisticated measures. Over the past few decades, mean amplitude (rather than peak amplitude) gradually became the standard.

The main reason why mean amplitude is usually thought to be superior to peak amplitude is that peak amplitude is easily influenced by noise. Indeed, as Clayson, Baldwin & Larson (2013) illustrated in their simulation study, increases in noise drastically affect the peak amplitude measure.

Crucially, this is true when the peak measure is extracted from the *original (observed) data* of single trials (and when taken from the

[3] It is also possible to calculate the Response Magnitude Index based on the GAM-fitted smooth values in the same time window, rather than the original values. We did this for the empirical example presented later in this paper. The results showed a very strong correlation between these two measures of $r = .96$, $p < .001$.

Table 1
The individual difference measures.

Measure		Unit	What does it quantify?
1.	Response Magnitude Index	Amplitude in μV	Size of response
2.	Modelled Area	Amplitude in μV * time in ms	Size of response
3.	Height Modelled Peak	Amplitude in μV	Size of response
4.	Normalized Modelled Peak	Unit of standard deviation (value >1: reliable peak, 0–1: large item variation)	Robustness of response (response stability)
5.	Modelled Area Median Latency	Time in ms	Timing of response
6.	Modelled Peak Latency	Time in ms	Timing of response

Note. Measure 1 is the traditional Response Magnitude Index based on the averaged original data, measures 2–6 are based on a GAM fitted smooth of the difference waveform.

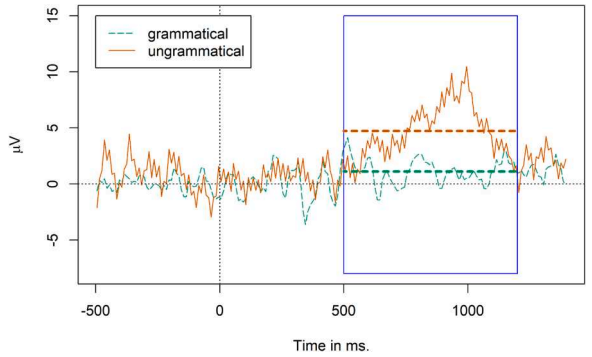

Fig. 1. Extracting the traditional Response Magnitude Index (Measure 1).
Note. This figure depicts the ERP signal averaged (over trials) for one participant. The green dashed line represents the averaged waveform for grammatically correct targets, and the orange solid line the averaged waveform for the ungrammatical targets. The blue rectangle depicts the time window over which the average amplitude is calculated (means are represented by the bold dashed horizontal lines). The amplitude difference between the two dashed lines is the Response Magnitude Index, which would equal around 3.6 μV in this example.

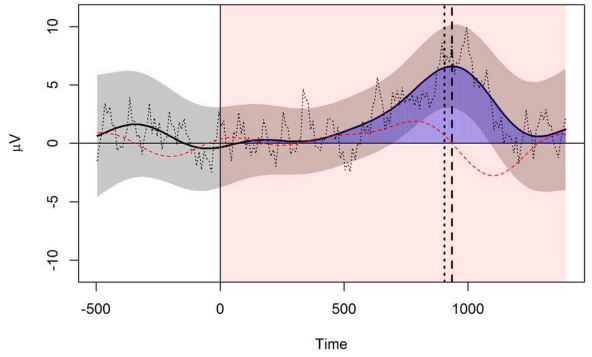

Fig. 2. Extracting the GAM-based measures (Measure 2–6).
Note. This figure depicts the GAM estimated ERP signal for all trials by one participant. The solid black line shows the GAM-smooth of the difference waveform (ungrammatical – grammatical). The dotted black signal shows the difference waveform of the averaged original data. The red shaded area marks the selected search window (as broad as possible to prevent a subjective bias of the researcher). The grey shaded area marks the 95% confidence bands of the GAM-smooth. The Modelled Area (*measure 2*) is marked in blue (in this example its size, μV*ms, would be 3048 – note that, as explained above, the smooth becomes zero for the grammatical condition). The point marking the Modelled Area Median Latency (*measure 5*) is shown as the dotted vertical line, here with a latency of 905 ms. The height and latency of the modelled peak are found using the derivative of the GAM smooth. The peak can be found where the derivative, shown as the red dashed line, crosses the x-axis (i.e., equals zero), representing the point at which the GAM-smooth stops increasing and starts decreasing. In this example, the Height of the Modelled Peak (*measure 3*) would be around 6.6 μV, and the Modelled Peak Latency (*measure 6*; shown by the vertical dashed line) around 937 ms. Finally, the Normalized Modelled Peak (*measure 4*) is found by dividing the height of the peak by 1.96 times the standard error of this height (in this example the value would be 1.9, which as the value is higher than 1 means the peak is robust).

averaged ERP waveform, there are other issues such as effects of latency jitter and bias by the number of trials; see Luck, 2014). Our approach differs in that we extract peak measures from the *modelled* (GAM) waveform. As explained in Section 2.1, the GAM appropriately deals with item variation, making the peak measures we extract from this waveform much less sensitive to noise and outliers, latency jitter and bias by the number of trials, as well as avoiding the necessity of selecting an arbitrary time window which may disadvantage some participants. Furthermore, a standard error of the height of a modelled peak can be calculated, allowing us to assess its robustness.

Before any peak measure can be extracted, two aspects need to be specified: a search window and the polarity of the peak of interest. As we described above, we recommend extracting the GAM-based measure from the modelled difference waveform, which should single out the component(s) of interest. This means that the search window can be sufficiently wide to capture individual (latency) differences. In the empirical example presented in Section 3 we use the full search window of 0 to 1400 ms, to demonstrate how we can study the P600 effect without imposing bias from an arbitrarily chosen (narrower) time window. By specifying the polarity, we determine whether we are interested in a positive component (such as the P600) or a negative component (such as the N400).

To find the peak in the (GAM-)modelled difference wave, we use the derivative[4] of the GAM function (plotted as the red dashed line in Fig. 2). In case we are looking for a positive peak, we find the measurement point for which the derivative equals zero, the previous point's derivate is positive (i.e., the slope is increasing before the peak) *and* the next point's derivative is negative (i.e., the slope is decreasing after the peak).[5]

For this measurement point we extract the Height of the Modelled Peak (*measure 3*), and the Modelled Peak Latency (*measure 6*). The Normalized Modelled Peak (*measure 4*) is calculated by dividing the height of the modelled peak by 1.96 multiplied by the standard error (SE). This normalisation step provides us with a measure of peak robustness. Effectively it indicates how many (95%) confidence bands above baseline the effect is. If the value is larger than 1, it means there is a reliable peak (i.e., the confidence band of the peak is not overlapping with the *x*-axis). If the value is between 0 and 1, then there is a large amount of item variation (i.e., the confidence band of the peak is overlapping with the *x*-axis), and thus the presence of the peak in this individual is less reliable.

In the following, we describe an empirical example, using data from a previously published experiment (Meulman et al, 2015), to demonstrate whether the GAM-based individual difference measures are indeed capable of capturing the effects they intend to measure, and how the measures may be applied in studies on L2 processing.

3. Empirical example: P600 paradigm

3.1. Dataset and analysis

We re-analysed data from a study on the processing of grammatical gender violations in native Slavic L2 speakers of German (Meulman et al, 2015). The benefit of using this dataset for the present demonstration is that it contains a relatively large sample (compared to most ERP studies) of N=66 L2 speakers and N=29 native speakers of German, and that the effects that are present in the data are known and have been reported in Meulman et al. (2015). Specifically, whereas native speakers consistently showed a P600 for gender violations, the researchers found a gradual change in linguistic processing for the L2 speakers that varied by age of acquisition of the L2, with L2 speakers that started learning at an earlier age showing a P600 and L2 speakers that started learning at a later age showing a posterior negativity. In other words, there is systematic variability across individuals in showing the P600 effect, which may be further investigated using the GAM-based individual difference measures. We confine ourselves to a very brief description of the participants, materials and procedures of the data that have been included in the current analysis below, and refer the reader to Meulman et al. (2015) for further details.

The current analysis includes data from a total of 95 participants: 66 highly proficient second language speakers of German (with a Russian or Polish L1 background) and 29 German native speakers. Participant characteristics are summarized in Table 2, and all participants were right-handed and did not have any cognitive, visual, or auditory impairments. L2 proficiency was assessed by means of a C-test (constructed by Schmid, 2011, available at https://languageattrition.org), which consisted of two texts containing gaps where parts of some words had been left out. The participants' task was to fill the gaps, and their percentage of correct answers was taken as their L2 proficiency score in the analysis.

In the EEG experiment, the participants listened to auditorily presented sentences while their electroencephalogram was recorded.

[4] The derivative, in mathematics, is the rate of change of a function with respect to a variable.

[5] A few other conditions need to be specified in order to select the Modelled Peak of interest. These conditions may differ depending on the setup of the experiment and/or component of interest. In the empirical example presented later in this paper, we 1) used the highest peak if multiple peaks were detected in the defined search window, and 2) took the highest point of the GAM smooth in the defined search window if no clear peak was detected: for all participants for which this was the case (21 L2 speakers and 4 native speakers) this meant that the highest point was at the end of the search and measurement window (1400 ms), therefore being our best proxy of the peak (which is assumed to occur after the search window). Of course, one may also choose to not extract any peak measures when no clear peak is detected, and simply distinguish individuals with a clear peak from those without a clear peak.

Table 2
Participant characteristics.

Characteristic or measure		L2 speakers (n = 66)	Native speakers (n = 29)
Age and exposure	Age at testing in years	28.9 (18–53)	37.8 (22–58)
	AoA in years[a]	17.7 (7–36)	–
	AoE in years[b]	15 (7–32)	–
	LoR in years[c]	11.3 (4–25)	–
	Sex	61F, 5M	19F, 10M
Proficiency measures:	C-test in %[d]	80.9 (51–95)	93.2 (86–98)
	Gender assignment in %[e]	93.3 (72.9–100)	99.9 (99–100)

[a] Age of acquisition (= age of arrival in the L2 country)
[b] Age of first exposure to the L2 (either in the L2 country or in a classroom setting outside of the L2 country)
[c] Length of residence in the L2 country
[d] Percentage of correct responses in the C-test
[e] Percentage of correct responses in the pen-and-paper gender assignment task

Some of the sentences contained violations in grammatical gender agreement (48 grammatical and 48 ungrammatical trials per participant).[6] An example sentence is given below (the critical target from which the ERP was measured is underlined and the * indicates the incorrect form of the determiner):

Nach der Schlägerei ist das/ ∗ der Auge des Angestellten von der Krankenschwester versorgt worden. (*After the fight the*$_{neut}$ / ∗*the*$_{masc}$ *eye of the worker was treated by the nurse.*)

After acquisition, the EEG data was re-referenced, filtered (<0.1Hz and >40Hz), segmented and baselined. Average ERPs (for the traditional approach) and GAM smooths (for the GAM-based approach) were obtained for trials free of artifacts without regard to behavioural responses. Specifically, ocular artefacts were corrected, but segments with other artefacts were removed. A central-posterior region of interest containing six electrodes (P3, Pz, P4, O1, Oz, O2) was used for analysis,[7] consistent with the location in which other studies on grammatical gender processing have found the (late stage) P600 effect to be most pronounced (Molinaro, Barber, & Carreiras, 2011). The data was down-sampled to 100 Hz for analysis.[8]

All analyses presented below were performed in R (version 4.2.0: R Core Team, 2022), and linear regression as well as generalized additive modelling analyses were performed using the mgcv R package (version 1.8.40: Wood, 2017; Wood, Goude, & Shaw, 2015). For reproducibility, the data, analysis and results are available as a paper package stored at the Open Science Framework repository (https://osf.io/zkd47/).

In order to isolate the ERP component of interest, we fitted our GAM on the difference waveform (i.e., in our case the difference between the ungrammatical and the grammatical condition) across the time-locked EEG signal per trial in the complete range of -500 to 1400 ms before/after the onset of the target word, using the model specification shown above (Section 2.1). These difference smooths are very helpful in showing the nature of the experimental effects and making the time course of the effects clear, by subtracting away everything except the one aspect that differs across conditions (Luck, 2014).

The traditional and GAM-based individual difference measures were extracted following the procedure described in Section 2.2 above. For the traditional Response Magnitude Index, a time window was selected of 500–1200 ms post stimulus onset, in line with what was used in the original Meulman et al. (2015) paper. This window, which is somewhat longer than is typical in P600 studies, was chosen based on visual inspection of the data to identify where the main effect lies, in order to create the optimal conditions for the Response Magnitude Index to perform well against the novel measures.[9] For the GAM-based measures, a search window from 0-1400 ms was specified, and the polarity was specified to 'positive', as we are interested in the P600 effect. The plots illustrating the GAM smooths and accompanying individual difference measures (similar to Fig. 2) for all participants are available in the paper package.

3.2. Comparison between the measures: distributions and correlations

Figs. 3 and 4 show the distribution of values for the various measures. Fig. 3 shows the measures of the size/robustness of the

[6] In total, each participant listened to 278 sentences, as in addition to the grammatical gender condition the experiment also contained sentences with verb agreement violations and a number of well-formed filler sentences. The verb agreement condition was not included in the current paper, since these violations consistently elicited a P600 effect throughout the L2 speakers as well as the native speakers, making this condition less interesting to use as an example for an investigation of individual differences.

[7] A GAM analysis would be able to deal with multiple electrodes, and this is certainly worth pursuing as scalp distribution provides another source of inter-participant variability. However, our current goal is to provide a clear demonstration of our measures in comparison the approach most usually applied, which is why we decided to focus on one region of interest.

[8] The data was down-sampled to 100 Hz to reduce processing time of the GAMs. Although this does result in loss of data, it should not affect the results of our investigation of the P600, which is a large, slow-going effect.

[9] To assess the effect of choosing a different time window for the Response Magnitude Index, we have also experimented with using a more commonly used time window ranging from 500 to 800 ms. See footnote 12 for those results.

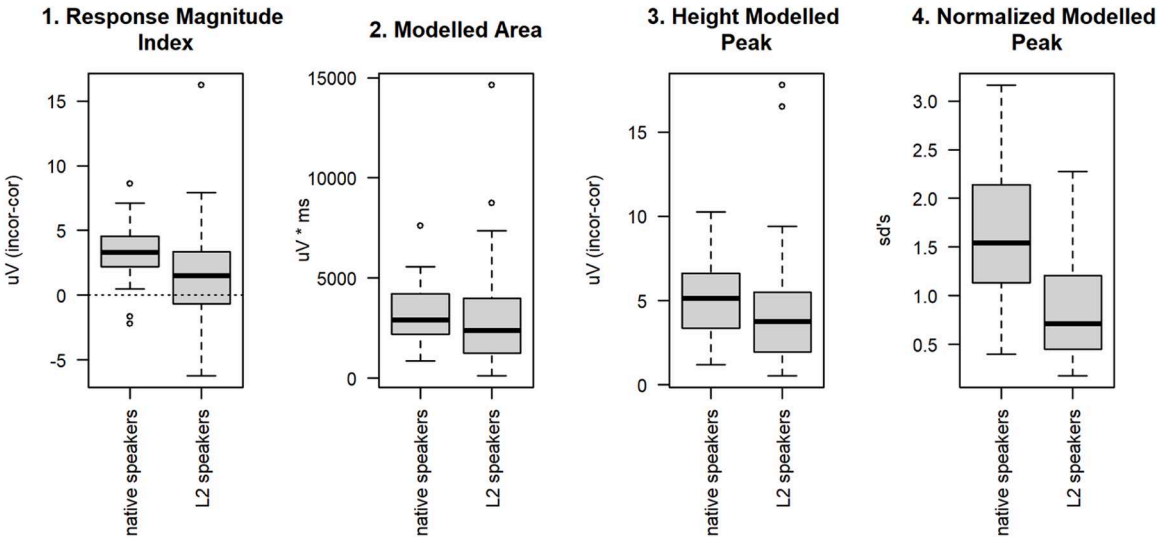

Fig. 3. Distribution of values for the amplitude measures.

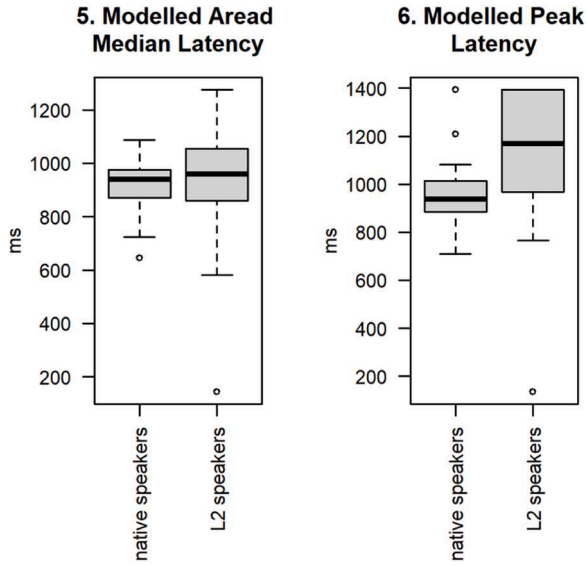

Fig. 4. Distribution of values for the latency measures.

response. Comparing the traditional Response Magnitude Index (*measure 1*), the Modelled Area (*measure 2*) and the Height of the Modelled Peak (*measure 3*), we see that all measures show a similar pattern, with higher amplitudes for native speakers compared to L2 speakers, and the L2 speaker group showing somewhat more variation than the native group. In the Normalized Modelled Peak (*measure 4*) we see that the distributions of the native and L2 speaker group are further apart, with the native speakers showing far more stable peaks (i.e., those with values of at least 1) than the L2 speakers.

Fig. 4 shows the distributions of the two latency measures. In the Modelled Area Median Latency (*measure 5*), the mean latencies in the L2 speaker group and the native group are similar, but the L2 speaker group shows more variability. In the Modelled Peak Latency (*measure 6*) latencies for the L2 speaker group are much higher than for the native group, and the variability among L2 speakers is also larger than among native speakers.

In nearly all of the plots presented in Fig.s 3 and 4 a few outliers are visible. In some cases, these datapoints are not likely to constitute a meaningful representation of the P600 effect. However, to be on the cautious side and not *a priori* assume measurement error, we retained all data points in the analysis.

The paired correlations between all measures are listed in Table 3. We see that correlations between the measures aimed at quantifying size of the response are strong: ranging from $r = .74$ for the Response Magnitude Index with the Height of the Modelled Peak, and $r = .86$ for the Modelled Area measure with the Height of the Modelled Peak, to $r = .92$ for the Response Magnitude Index

Table 3
Pairwise correlations between the measures.

	Response Magnitude Index	Modelled Area	Height Modelled Peak	Normalized Modelled Peak	Modelled Area Median Latency
Response Magnitude Index					
Modelled Area	0.92***				
Height Modelled Peak	0.74***	0.86***			
Normalized Modelled Peak	0.59***	0.47***	0.52***		
Modelled Area Median Latency	-0.06	0.06	0.27*	0.10	
Modelled Peak Latency	-0.20	-0.02	0.00	-0.35**	0.62***

*** *Note.* $p < .001$
** $p < .01$
* $p < .05$.

with the Modelled Area measure. The Normalized Modelled Peak correlates moderately with the other amplitude measures: $r = .47$, $r = .52$ and $r = .59$ for the Modelled Area, Height of the Modelled Peak and Response Magnitude Index, respectively.

The latency measures correlate $r = .62$ with each other. Furthermore, there are weak correlations between the Modelled Area Median Latency and the Height of the Modelled Peak ($r = .27$), and between the Modelled Peak Latency and the Normalized Modelled Peak ($r = -.35$). There are no other significant ($p < .05$) correlations between the latency and amplitude measures (*r*s between -.20 and .10).

From these correlations we can conclude that the three measures aimed at quantifying the size of the response (or the amount of neural activity) are largely measuring the same construct. The two measures of latency also show a correlation, but somewhat less strong. Furthermore, we observe that a high magnitude or peak of the response does not always coincide with a robust peak. Finally, the latency of the response seems to be somewhat independent of the size, although there is a hint that later peaks go together with less reliable peaks due to the low negative correlation.

3.3. Systematic individual variability: predictors of group and proficiency

In order to determine the suitability of these measures for predicting systematic variability at both group and individual levels, we first fitted a binomial linear regression model with *group* (nonnative = 1 vs. native = 0) as the dependent variable,[10] assessing the impact of the relevant measure as the predictor. We fitted separate models for each measure in order to estimate which of them had greater explanatory power as a predictor of group. The results of this analysis (see Table 4) show that the Normalized Modelled Peak is by far the strongest predictor of the group effect. The model including this predictor explains 25.60% of the deviance, whereas the other measures only explain between 0.05% and 7.07%. The negative coefficient for the Normalized Modelled Peak measure means that a more robust (i.e., reliable) peak is less likely to be associated with an L2 speaker than with a native speaker ($\beta = -2.181$, SE = 0.564, $p < .001$).

The only other significant predictor of group is Modelled Peak Latency, with a higher value meaning that a later peak is more likely to associated with an L2 speaker ($\beta = 0.003$, SE = 0.001, $p = .017$) than a native speaker. This model, with Modelled Peak Latency as predictor of group, explains about 7.07% of the deviance.

The Response Magnitude Index, the Modelled Area measure, the Height of the Modelled Peak and the Modelled Area Median Latency only explain about 2.59%, 0.05%, 0.70% and 0.18% of the deviance in group, respectively, and none of them are significant predictors of group (all *p*s $> .1$).

Similarly, we fitted a binomial linear regression model for each of the measures with *proficiency*, as indexed by the C-test (answers correct vs. answers incorrect, with scores ranging from 51 to 95 percent correct), as the dependent variable. Again, we assessed the impact of the relevant measure as predictor in these models, to see which had the greatest explanatory power. The results of this analysis, which focusses on the L2 speaker group only (as almost all monolinguals perform at ceiling), are presented in Table 5 and again show that the Normalized Modelled Peak is the strongest predictor of proficiency: the model with this predictor explains 4.12% of the deviance. A higher value for the Normalized Modelled Peak measure predicts a higher score on the proficiency index ($\beta = 0.348$, SE = 0.090, $p < .001$), which means that L2 speakers with a higher proficiency have more robust peaks than L2 speakers with a lower proficiency.

Here, the only other significant predictor of proficiency is the traditional Response Magnitude Index, with a higher value predicting a higher score on the proficiency index ($\beta = 0.042$, SE = 0.015, $p = .005$), i.e., L2 speakers with higher proficiency have a higher average response magnitude than L2 speakers with lower proficiency. This model, with Response Magnitude Index as predictor of proficiency, explained 2.30% of the deviance.

[10] Note that in this analysis we use group and proficiency as dependent variables, whereas typically, in an investigation of group or proficiency effects, these would be the independent rather than the dependent variables in the model. However, the objective of our analysis was not to investigate group or proficiency effects as such, but rather to evaluate how well the different ERP measures are able to predict group membership and proficiency.

Table 4
Results for the Predictors of group (L2 Speakers vs. Native speakers).

Predictor	Estimate	Std. error	z-value	*p*-value	Deviance explained
Response Magnitude Index	-0.156	0.106	-1.471	.141	2.59%
Modelled Area	<-0.001	<0.001	-0.221	.825	0.05%
Height Modelled Peak	-0.063	0.078	-0.809	.418	0.70%
Normalized Modelled Peak	-2.181	0.564	-3.867	<.001 ***	25.60%
Modelled Area Median Latency	<0.001	0.002	0.412	.680	0.18%
Modelled Peak Latency	0.003	0.001	2.395	.017 *	7.07%

Note. Each row of the table represents a separate binomial linear regression model with *IsL2speaker* as the dependent variable, assessing the impact of the relevant predictor. Each model was fitted on the basis of 72 data points.[111]

Table 5
Results for the predictors of proficiency (within the L2 speakers)

Predictor	Estimate	Std. error	z-value	*p*-value	Deviance explained
Response Magnitude Index	0.042	0.015	2.841	.005**	2.30%
Modelled Area	<0.001	<0.001	0.093	.926	<0.01%
Height Modelled Peak	-0.008	0.011	-0.756	.449	0.15%
Normalized Modelled Peak	0.348	0.090	3.846	<.001***	4.12%
Modelled Area Median Latency	<0.001	<0.001	1.383	.167	0.50%
Modelled Peak Latency	<-0.001	<0.001	-1.557	.120	0.66%

Note. Each row of the table represents a separate linear regression model with proficiency (as measures by a standard C-test) as the dependent variable, assessing the impact of the relevant predictor. Each model was fitted on 46 data points.

The Modelled Area measure, the Height of the Modelled Peak and the two latency measures only explained between 0% and 0.66% of the deviance in proficiency and none of them were significant predictors (all *p*s $> .1$).

From the results presented above, we can conclude that the Normalized Modelled Peak is the best predictor of both group and proficiency. The Modelled Peak Latency is also able to distinguish between L2 speakers and native speakers at the group level, and the traditional Response Magnitude Index is also able to distinguish between more and less proficient L2 speakers.[12] None of the other measures, quantifying size or latency of the ERP response, reached significance as a predictor.

We would like to end this section by making two remarks about the analysis we just presented. First, it is worth noting that the mere presence of a peak (i.e., whether we can detect a peak in the GAM function of the waveform, coded as a binary true/false variable) was also a strong and significant predictor of the group effect ($\beta = -1.530$, SE $= 0.619$, $p < .05$; deviance explained $= 7.68\%$), and of the proficiency effect ($\beta = 0.438$, SE $= 0.079$, $p < .001$; deviance explained $= 8.39\%$). However, as the Normalized Modelled Peak provides additional information about the robustness of the peak compared to the mere presence of the peak, we only included the latter in our overview.

Second, these findings are robust, as the Normalized Modelled Peak remained the strongest predictor of both group membership and proficiency irrespective of differently selected time and search windows.

4. Discussion

This paper set out to create novel individual difference measures, in order to provide more sensitive and complete participant-specific measures for quantifying the size, latency and stability of ERP components (e.g., the P600). These types of measures are needed as a tool for researchers studying bilingual development and ultimate attainment. In this line of research, individual differences are an important topic of investigation. We have only just begun to discover the factors that modulate between-participant variation in ERP responses, in part due to having previously lacked adequate tools to do so. The traditionally-used Response Magnitude Index (Tanner, Mclaughlin, Herschensohn & Osterhout, 2013), suffers from a number of drawbacks. By averaging (over items and across a time window), it may not constitute the best representation of the size of an individual's response, since outliers and latency differences may distort these averages and preclude studying latency effects on an individual basis. Furthermore, this measure only provides insight into the (average) *size* of each participant's ERP component, but it does not capture the *robustness* of an L2 speaker's ERP response.

Our approach thus uses generalized additive modelling (GAM) to *model* the ERP waveform. As a single-trial regression-based analysis, this approach is much less sensitive to outliers. From the modelled waveform, we extracted several measures without the need to specify a particular time window. The Modelled Area measure took the geometric area under the curve as a measure of size of the

[12] When instead of a time window of 500 – 1200 ms., we use a more restricted P600 time window of 500 – 800 ms for the Response Magnitude Index, the performance of the measure to reveal individual proficiency differences is diminished and becomes non-significant (deviance explained: 0.89%, $p = 0.066$). With respect to being able to detect group differences, the measure remains non-significant (deviance explained: 3.70%, $p = 0.074$).

ERP response, and the Modelled Area Median Latency as a measure of timing. The modelled peak was extracted to calculate measures of size, latency and robustness of the ERP response: the Height of the Modelled Peak, the Modelled Peak Latency, and the Normalized Modelled Peak.

In an empirical example, we used data from a previously published experiment (Meulman et al, 2015) with a relatively large number of L2 speaker and native participants in a P600 paradigm investigating grammatical gender agreement processing. We investigated how well our GAM-based individual difference measures were able to capture the individual variation and how the measures compared to the traditional Response Magnitude Index, and to each other, in this respect.

We showed that, in our sample, the three amplitude measures (traditional Response Magnitude Index, Modelled Area and Height of the Modelled Peak) correlated strongly with each other, whereas the Normalized Modelled Peak correlated moderately with these amplitude measures, which indicates that the measure of robustness captures a partly different construct than the measures of size, and that a large ERP response does not always coincide with a stable and robust ERP response. Furthermore, we showed that the Normalized Modelled Peak was the measure that was best able to discriminate between L2 speakers and native speakers (in addition to the latency of this peak), and between different levels of L2 proficiency (as measures by a C-test) in L2 speakers. The traditional Response Magnitude Index was also able to distinguish between more and less proficient L2 speakers (although to a lesser extent than the Normalized Modelled Peak), but the other amplitude measures did not capture the group and proficiency effects.

The two measures of latency (Modelled Area Median Latency and Modelled Peak Latency) showed a moderately strong correlation between each other. Latency of the response seems to be somewhat independent of the height of the response, although later peaks sometimes tend to go together with less reliable peaks. Modelled Peak Latency is a relatively good predictor for explaining differences between the native and the L2 speaker group (although to a lesser degree than the Normalized Modelled Peak). But neither Modelled Area Median Latency nor Modelled Peak Latency are good at distinguishing proficiency differences between L2 individuals. Plot 6 of Fig. 4 gives us some indication of why that may be. We see that in the dataset in our empirical example, P600 peaks are rather late (even for the native speakers, but particularly for the L2 speakers), which may be specific to this dataset, for example, due to the fact that an experimental design with auditory sentence presentation was used. As a consequence, there is a truncated range (limited by the end of the measurement window at 1400 ms) for the L2 group.

The present results are important for at least two reasons. First, we found that the height of the ERP peak did not discriminate well between L2 speakers and native speakers, or more and less proficient L2 speakers. Instead, the *robustness* (i.e., reliability) of this peak did prove to be a strong, distinctive measure. This is in line with current views on the P600 effect reflecting a process that either occurs or does not occur at a critical word. The actual size of an individual's ERP component is dependent on many factors, and variation is found even in monolinguals (Pakulak & Neville, 2010; Beatty-Martínez et al. 2021; Tanner, 2019). We showed that size and robustness of the modelled peak at least partly reflect a different construct (with a correlation of .52, i.e. only 27% overlap), and that the Normalized Modelled Peak (i.e., peak robustness) captured a much larger part of the individual variability between L2 speakers and native speakers (25.6% versus 0.7% for the Height of the Modelled Peak) and in the proficiency score differences between L2 speakers (4.1% versus 0.2%, respectively). We conclude that the Normalized Modelled Peak reflects the stability of an individual's P600 response, which, at least in this dataset, is modified by whether the individual is an L2 speaker or a native speaker, and by an L2 speaker's level of L2 proficiency. These results suggest that this measure may be a suitable tool for future studies on bilingual development. While the traditional measure (Response Magnitude Index) was able to distinguish between different levels of L2 proficiency, and to a limited extent also distinguish between native speakers and L2 speakers, the Normalized Modelled Peak was much more sensitive, particularly for the latter task.

Second, whereas latency effects have traditionally only been investigated in grand average waveforms of L2 speakers, we believe this is the first study to investigate differences in timing of ERP responses between L2 speakers at the individual level. Although in our dataset the GAM-based latency measures were unable to distinguish more proficient from less proficient L2 speakers, the Modelled Peak Latency was a strong predictor for determining whether an individual belonged to the L2 speaker or the native speaker group. This suggests that response delays may be an inherent factor of bilingual processing, due to the higher cognitive load incurred by processing more than one language (as discussed above) as such, rather than indexing different proficiency levels. These results suggest that this latency measure may therefore also be a suitable tool for studies on bilingual language processing.

Nevertheless, further research is needed to evaluate these claims. We included several measures of response amplitude and latency in this paper, as an exploration into what could be a good measure to quantify individual differences in L2 speakers' ERP responses. Based on only one study it is impossible to establish the performance of each of the measures we presented and conclusively determine which measures are (most) useful. We therefore hope other researchers will apply these measures to their datasets, further assessing the validity and useability of the measures.[13]

Even though the Modelled Area measure was included to quantify the size of the response without being affected by latency

[11] Since our investigation focused on P600 effects, participants not showing any effect above baseline in the search window were automatically left out here, since (positive polarity) Modelled Area and Modelled Peak measures could not be calculated for these individuals (these were 20 L2 speakers and 3 native speakers, who showed a negativity instead of a positivity). A separate analysis using GAM-based individual difference measures could be conducted focusing on negative polarity effects, which, together with the analysis we presented here, gives a complete picture of the individual variation present in this dataset. However, our current goal is to provide a clear demonstration of our new measures, which is why we decided to keep the analysis simple and focus on one polarity.

[13] All necessary functions to conduct these analyses are available in a paper package stored at the Open Science Framework repository (https://osf.io/zkd47/) to facilitate their uptake.

differences between individuals, this measure did not outperform the traditional Response Magnitude Index in our empirical example. The explanation for this likely lies in the fact that in our data set any P600 effects found did not suffer from excessive cancellation from previous or subsequent waves in the opposite direction. In other studies, however, such opposite waves may be present. In that case, in order to calculate the Response Magnitude Index, the researchers may have to resort to selecting a narrower window on the basis of the observed time course of the effect in the grand average waveform, introducing bias against individuals showing a different time course (and the narrower window also may be less sensitive, as we found when evaluating a shorter time window for the Response Magnitude Index; see footnote 12). In such a case, assessing the suitability of the Modelled Area measure over the Response Magnitude Index would be useful.

Furthermore, we have some indication that it may be worthwhile in future research to investigate the significance of the binary absence/presence of a peak in an individual's GAM waveform. In our empirical example, the presence of a true peak (as opposed to an increasing signal across the whole search window) in the GAM smooth of the waveform was also a strong and significant predictor of the group effect and of the proficiency effect (explained deviance 7.7% and 8.4%, respectively – in the latter case it therefore even outperformed the Normalized Modelled Peak). In further exploratory analysis (not presented in this paper, but available in the paper package) we found that when looking at the subset of the data with only true peaks both Modelled Area Median Latency and Modelled Peak Latency are excellent predictors of proficiency differences between L2 speakers (both explaining around 15% of the deviance). The reason for this difference is that, in the absence of a true peak, these latencies are very high, as they simply represent the end of the search window. However, the subset of data containing true peaks consists of only 47 out of 72 data points, so some caution in interpreting this result is necessary. Future research is needed to determine the significance of the presence or absence of a true peak in the modelled waveform, before any strong conclusions can be drawn from this analysis. Nevertheless, our observations do suggest that it is worthwhile to further investigate which factors modulate the latency of the modelled peak in L2 populations.

To develop a full picture of inter-participant variability, future studies should furthermore incorporate other sources of variation across all dimensions of the ERP waveform. For example, our approach may be extended to investigate inter-participant variability in scalp distribution, and as a GAM analysis would be able to deal with multiple electrodes sites individually, this would be the recommended technique.

In summary, we proposed a new method to extract a set of individual difference measures from the ERP waveforms, modelled using GAMs. The advantages of our technique, compared to the traditional approach of extracting the Response Magnitude Index, are that the GAM-based individual difference measures are less dependent on pre-specified time-windows (therefore latency differences between individuals are not penalized), are less sensitive to outliers, and allow for the investigation of latency differences between individuals. Our method can be applied to any ERP component (and also to any group of participants, i.e., the technique is not limited to investigate ERP effects in L2 speakers), as long as the component of interest can be isolated via determining the difference smooth. A disadvantage of the present method is that it does (in principle) not distinguish between participants who show a clear peak in the ERP effect from those who do not. Nevertheless, running the analysis costs relatively little effort, as all R code has been made available (https://osf.io/zkd47). The technique does not require more data (either subjects or items) than is usual for ERP studies. Of course, using fewer items will make it harder to detect robust peaks (due to larger standard errors). The algorithm is relatively fast: processing one participant takes only a few seconds.

Of the presented GAM-based individual difference measures, the Normalized Modelled Peak in particular appears to be a sensitive measure. Specifically, it appears to outperform the traditionally used Response Magnitude Index in distinguishing native and non-native speakers or more proficient from less proficient L2 speakers. The former is also true for the Modelled Peak Latency, which can be used to study individual differences in timing of the ERP response.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Monika S. Schmid reports financial support was provided by Dutch Research Council (NWO).

Acknowledgements

This research was supported by the Netherlands Organization for Scientific Research (NWO) under grant 016.104.602, awarded to M.S.S. We thank Jacolien van Rij and two anonymous reviewers for discussion and comments on previous versions of this paper.

References

Abugaber, D., Finestrat, I., Luque, A., & Morgan-Short, K. (2023). Generalized additive mixed modeling of EEG supports dual-route accounts of morphosyntax in suggesting no word frequency effects on processing of regular grammatical forms. *Journal of Neurolinguistics, 67*, Article 101137. https://doi.org/10.1016/j.jneuroling.2023.101137

Alemán Bañón, J., Miller, D., & Rothman, J (2017). Morphological variability in second language learners: An examination of electrophysiological and production data. *Journal of Experimental Psychology: Learning Memory and Cognition, 43*, 1509–1536. https://doi.org/10.1037/xlm0000394. https://psycnet.apa.org/

Alemán Bañón, J., Fiorentino, R., & Gabriele, A (2018). Using event-related potentials to track morphosyntactic development in second language learners: The processing of number and gender agreement in Spanish. *PLoS One, 13*(7), Article e0200791. https://doi.org/10.1371/journal.pone.0200791

Beatty-Martínez, A. L., Bruni, M. R., Bajo, M. T., & Dussias, P. E. (2021). Brain potentials reveal differential processing of masculine and feminine grammatical gender in native Spanish speakers. *Psychophysiology, 58*(3), e13737. https://doi.org/10.1111/psyp.13737

Bialystok, E. (2009). Bilingualism: The good, the bad, and the indifferent. *Bilingualism: Language and Cognition, 12*(1), 3–11. https://doi.org/10.1017/S1366728908003477

Bice, K., & Kroll, J. F. (2021). Grammatical processing in two languages: How individual differences in language experience and cognitive abilities shape comprehension in heritage bilinguals. *Journal of Neurolinguistics, 58*, Article 100963. https://doi.org/10.1016/j.jneuroling.2020.100963

Bond, K., Gabriele, A., Fiorentino, R., & Alemán Bañón, J (2011). Individual differences and the role of the L1 in L2 processing : An ERP investigation. In *Proceedings of the 11th Generative Approaches to Second Language Acquisition Conference* (pp. 17–29).

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science, 41*(6), 1318–1352. https://doi.org/10.1111/cogs.12461. *Suppl.*

Carrasco-Ortíz, H., Herrera, A. V., Jackson-Maldonado, D., Ramírez, G. N. A., Pereyra, J. S., & Wicha, N. Y. (2017). The role of language similarity in processing second language morphosyntax: Evidence from ERPs. *International Journal of Psychophysiology, 117*, 91–110. https://doi.org/10.1016/j.ijpsycho.2017.04.008

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology, 50*(2), 174–186. https://doi.org/10.1111/psyp.12001

Donchin, E., & Heffley, E. F., III (1978). Multivariate analysis of event-related potential data: A tutorial review. In D. Otto (Ed.), *Multidisciplinary Perspectives in Event-Related Brain Potential Research* (pp. 555–572). Washington, DC: U.S: Government Printing Office.

Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language, 66*(1), 226–248. https://doi.org/10.1016/j.jml.2011.07.007

Green, D. W. (2011). Language control in different contexts: The behavioral ecology of bilingual speakers. *Frontiers in Psychology, 2*, 103. https://doi.org/10.3389/fpsyg.2011.00103

Grey, S., Tanner, D., & van Hell, J. G. (2017). How right is left? Handedness modulates neural responses during morphosyntactic processing. *Brain Research, 1669*, 27–43. https://doi.org/10.1016/j.brainres.2017.05.024

Grey, S. (2022). Variability in native and nonnative language: An ERP study of semantic and grammar processing. *Studies in Second Language Acquisition, 45*, 137–166. https://doi.org/10.1017/S0272263122000055

Hansen, J. C., & Hillyard, S. A. (1980). Endogeneous brain potentials associated with selective auditory attention. *Electroencephalography and clinical neurophysiology, 49*(3-4), 277–290. https://doi.org/10.1016/0013-4694(80)90222-9

Hopp, H. (2010). Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua, 120*(4), 901–931. https://doi.org/10.1016/j.lingua.2009.06.004

Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience, 15*(1), 98–110. https://doi.org/10.1162/089892903321107855

Kiesel, A., Miller, J., Jolicœur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology, 45*(2), 250–274. https://doi.org/10.1111/j.1469-8986.2007.00618.x

Kim, A. E., Oines, L., & Miyake, A. (2018). Individual differences in verbal working memory underlie a tradeoff between semantic and structural processing difficulty during language comprehension: An ERP investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(3), 406–420. https://doi.org/10.1037/xlm0000457. https://psycnet.apa.org/

Kos, M., Vosse, T. G., Van Den Brink, D., & Hagoort, P. (2010). About edible restaurants: Conflicts between syntax and semantics as revealed by ERPs. *Frontiers in Psychology, 1*(1–11), Article 21833277. https://doi.org/10.3389/fpsyg.2010.00222. pmid.

Kotz, S. A., Holcomb, P. J., & Osterhout, L. (2008). ERPs reveal comparable syntactic sentence processing in native and non-native readers of English. *Acta Psychologica, 128*(3), 514–527. https://doi.org/10.1016/j.actpsy.2007.10.003

Kotz, S. A. (2009). A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain and Language, 109*(2–3), 68–74. https://doi.org/10.1016/j.bandl.2008.06.002

Leckey, M., & Federmeier, K. D. (2019). The P3b and P600(s): Positive contributions to language comprehension. *Psychophysiology, 57*(7), e13351. https://doi.org/10.1111/psyp.13351

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't): How to get significant effects. *Psychophysiology, 54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Luck, S. J. (2014). *An introduction to the event-related potential technique* (Second Edition). Cambridge, Massachusetts: MIT press.

Meulman, N., Stowe, L. A., Sprenger, S. A., Bresser, M., & Schmid, M. S. (2014). An ERP study on L2 syntax processing: When do learners fail? *Frontiers in Psychology, 5*, 1072. https://doi.org/10.3389/fpsyg.2014.01072

Meulman, N., Wieling, M., Sprenger, S. A., Stowe, L. A., & Schmid, M. S. (2015). Age effects in L2 grammar processing as revealed by ERPs and how (not) to study them. *PLoS One, 10*(12), Article e0143328. https://doi.org/10.1371/journal.pone.0143328

Molinaro, N., Barber, H. A., & Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex, 47*(8), 908–930. https://doi.org/10.1016/j.cortex.2011.02.019

Morgan-Short, K., Sanz, C., Steinhauer, K., & Ullman, M. T. (2010). Second language acquisition of gender agreement in explicit and implicit training conditions: An event-related potential study. *Language Learning, 60*(1), 154–193. https://doi.org/10.1111/j.1467-9922.2009.00554.x

Morgan-Short, K. (2014). Electrophysiological approaches to understanding second language acquisition: A field reaching its potential. *Annual Review of Applied Linguistics, 34*, 15–36. https://doi.org/10.1017/S026719051400004X

Osterhout, L. (1997). On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences. *Brain and Language, 59*, 494–522. https://doi.org/10.1006/brln.1997.1793

Pélissier, M. (2020). Comparing ERPs between native speakers and second language learners: Dealing with individual variability. In A. Edmonds, P. Leclercq, & A. Gudmestad (Eds.), *Interpreting language-learning data* (pp. 39–69). Language Science Press. https://doi.org/10.5281/zenodo.4032282.

Pakulak, E., & Neville, H. J. (2010). Proficiency differences in syntactic processing of monolingual native speakers indexed by event-related potentials. *Journal of Cognitive Neuroscience, 22*, 2728–2744. https://doi.org/10.1162/jocn.2009.21393

Payne, B. R., Ng, S., Shantz, K., & Federmeier, K. D. (2020). Event-related brain potentials in multilingual language processing: The N's and P's. In *Psychology of Learning and Motivation, 72* pp. 75–118). https://doi.org/10.1016/bs.plm.2020.03.003

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Rossi, S., Gugler, M. F., Friederici, A. D., & Hahne, A. (2006). The impact of proficiency on syntactic second-language processing of German and Italian: Evidence from event-related potentials. *Journal of Cognitive Neuroscience, 18*(12), 2030–2048. https://doi.org/10.1162/jocn.2006.18.12.2030

Sabourin, L., & Stowe, L. A. (2008). Second language processing: when are first and second languages processed similarly? *Second Language Research, 24*(3), 397–430. https://doi.org/10.1177/0267658308090186

Sassenhagen, J., & Fiebach, C. J. (2019). Finding the P3 in the P600: Decoding shared neural mechanisms of responses to syntactic violations and oddball targets. *NeuroImage, 200*, 425–436. https://doi.org/10.1016/j.neuroimage.2019.06.048

Schmid, M. S., & Köpke, B. (2017). The relevance of first language attrition to theories of bilingual development. *Linguistic Approaches to Bilingualism, 7*(6), 637–667. https://doi.org/10.1075/lab.17058.sch

Schmid, M. S. (2011). *Language attrition*. Cambridge: University Press.

Smith, N. J., & Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology, 52*(2), 157–168.

Smith, N. J., & Kutas, M. (2015b). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology, 52*(2), 169–181. https://doi.org/10.1111/psyp.12317

Steinhauer, K. (2014). Event-related potentials (ERPs) in second language research: A brief introduction to the technique, a selected review, and an invitation to reconsider critical periods in L2. *Applied Linguistics, 35*(4), 393–417. https://doi.org/10.1093/applin/amu028

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia, 56*, 289–301. https://doi.org/10.1016/j.neuropsychologia.2014.02.002

Tanner, D., Mclaughlin, J., Herschensohn, J., & Osterhout, L. (2013). Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingualism: Language and Cognition, 16*(2), 367–382. https://doi.org/10.1017/S1366728912000302

Tanner, D., Inoue, K., & Osterhout, L. (2014). Brain-based individual differences in online L2 grammatical comprehension. *Bilingualism: Language and Cognition, 17*(2), 277–293. https://doi.org/10.1017/S1366728913000370

Tanner, D., Goldshtein, M., & Weissman, B. (2018). Individual differences in the real-time neural dynamics of language comprehension. *Psychology of learning and motivation, 68*, 299–335. https://doi.org/10.1016/bs.plm.2018.08.007

Tanner, D. (2019). Robust neurocognitive individual differences in grammatical agreement processing: A latent variable approach. *Cortex, 111*, 210–237. https://doi.org/10.1016/j.cortex.2018.10.011

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics, 70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wood, S. N., Goude, Y., & Shaw, S. (2015). Generalized additive models for large datasets. *Journal of the Royal Statistical Society: Series C Applied Statistics, 64*(1), 139–155. https://doi.org/10.1111/rssc.12068

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd Edn). Boca Raton: Chapman and Hall/CRC.