In press at Journal of Phonetics

Title: On visualizing phonetic data from repeated measures experiments with multiple random effects

Stephen Politzer-Ahles[a*], Page Piccinini[b]

*Corresponding author: stephen.politzerahles@polyu.edu.hk, Tel: +852 2766 7891
[a] The Hong Kong Polytechnic University, Hong Kong
[b] Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL* Research University, 75005, Paris, France

Abstract

In recent years, phonetic sciences has hosted several debates about how best to statistically analyze data. The main discussion has been about moving away from analysis of variances (ANOVAs) to linear mixed effects models. Mixed models have the advantage both of allowing for including all data points produced by a participant (instead of computing means for each participant) and accounting for both by-participant and by-item variance. However, plotting of data has not always followed this trend. Often researchers will plot participant means and standard error (as based on the number of participants), which while potentially representative of the data used for an ANOVA, does not match the data used for a mixed effects model. The present paper discusses the shortcomings of traditional data visualization practices, solutions to these shortcomings that have been discussed in recent years, and the special challenges that come with trying to extend these solutions to phonetic data with crossed (within-participant and within-item) designs. For each of the problems discussed, we provide examples with simulated data to demonstrate how different plotting techniques can correctly, or incorrectly, represent the underlying structure of data. Ultimately we conclude that there is no single type of plot that can show everything one needs to know about this type of data, and we advocate for an approach that involves using different types of plots throughout data analysis, and making data publicly available.

Keywords

data visualization, repeated measures, random effects, linear mixed effects models

In press at Journal of Phonetics

On Visualizing Phonetic Data from Repeated Measures Experiments with Multiple Random Effects

1 Introduction

The last ten years have seen many advances in statistical analyses in phonetic sciences. In inferential statistics, researchers have been using more advanced models such as linear mixed effects models (LMEMs) to test experimental hypotheses. These models give researchers the flexibility to account for several kinds of variances in the data. Predictive analytics has also become increasingly popular in phonetics, with researchers using predictive models to do things such as classify or cluster sounds given a set of phonetic features. Indeed, this special issue looks at the rise of different statistical methodologies in the field, trying to better understand which can potentially be most useful to phoneticians.

While most discussion has focused on statistical analyses, another aspect of data analysis that has begun to receive more attention is data visualization. The relationship between statistics and visualization is interesting, as the choice of a statistical test has often guided the method of visualization. For example, when using analyses of variance (ANOVAs), researchers often find the mean of a given dependent variable and then plot a representation of that mean, with a standard error based on the standard deviation, the number of participants, and the *t* distribution. However, this method can often obscure important information about the data, information that should inform the model used for statistical analysis.

Recently there has also been increasing awareness of the importance of considering variation when making statistical conclusions, and the value of visualizing the data rather than simply relying on dichotomous judgments (i.e., "significant" or "not significant") based on inferential statistics. Regardless of whether a pattern is significant, it is important to be aware of things like how many participants show the pattern, particularly when making conclusions about the practical or psychological importance of a finding. Accordingly, recent years have seen the publication of several papers with valuable exhortations and recommendations about how to improve our visualization practices to show how reliable (or un-reliable) results are across participants (Rousselet et al., 2016; Weissgerber et al., 2015). However, data from phonetics experiments (as well as other kinds of experimental psychology data, such as psycholinguistic data) often raise another problem, that of items or stimuli. In phonetics one usually wants to make general conclusions about some phenomenon—how something is realized in a given language, in a given speech context, or in a given population, etc. This requires not only generalizing beyond the participants who took part in the experiment, but also generalizing beyond the specific stimuli – words, sentences, contexts, etc. – that were used in the experiment. The need for statistical methods that allow for inferences beyond the items tested has been known for over 40 years (Clark, 1973), and in the last decade LMEMs have emerged as a popular and powerful technique to facilitate inferences both beyond the participants tested and beyond the items used (Baayen et al., 2008; Chang & Lane, 2016; Judd et al., 2012). Are there also visualization techniques to facilitate such inferences? Just as Rousselet and colleagues (2016) and Weissgerber and colleagues (2015) recommend that simple *t*-tests or ANOVAs should be supplemented by visualizations showing how a pattern varies across participants, it is also important for LMEMs to be supplemented by visualizations showing how a pattern varies across both participants and items (and whatever other relevant repeated-measures factors there are). Here we take up the question of whether or not this is possible.

In this paper we summarize some arguments that have been made to advocate for better data visualization practices, and go on to illustrate why it is challenging to carry these practices out effectively when it comes to common experimental designs in phonetics and psycholinguistics. After illustrating why it is not possible to simply plot all the data, we review some potential solutions and their limitations. We end by giving some suggestions for types of visualizations that show the important aspects of the data as much as possible, discussing the importance of using different types of visualizations for different purposes at different stages of data analysis and presentation, and emphasizing the importance of data availability. Through these examples we hope to illustrate that data visualization for designs typical in phonetics – experiments with repeated measures for both participants and items – is challenging and has no one-size-fits-all solution, but requires an awareness of the advantages and disadvantages of each visualization technique for each stage of data analysis.

2 Problems in data visualization

Data visualizations serve many different purposes, such as aiding steps of data analysis (like identification of outliers), informing statistical inferences, and communicating patterns of results to others. In this section we will focus on challenges in making statistical inferences from plots; in section 3 we will discuss other relevant functions of data visualizations.

2.1 Capturing the data distribution

A necessary first step in any analysis is to have a sense of your data's distribution. Lacking a full understanding of the distribution of your data can have implications for both data visualization and data analysis. In phonetic sciences it is common for data visualizations to show a measure of central tendency (e.g., a location parameter such as the mean) and a measure of variance or precision (e.g., error bars representing a scale parameter like standard deviation, or standard error). It has long been known, however, that such plots hide potentially important information about the shape of a distribution (e.g., Anscombe, 1973).

For example, imagine you have two conditions, and both conditions have the same mean and standard deviation, suggesting that they are the same. However, on closer inspection it turns out that the two conditions have very different distributions (e.g., normal and log), and as a result cannot be considered the same. See the example in Figure 1 showing how the same data can look different depending on how it is plotted (note that the code for this and all other plots in this article is available at https://osf.io/pm82v/). In the bar plot the two datasets look the same in terms of their means and standard deviations, but boxplots and scatter plots make it clear that they have different distributions; histograms would as well.

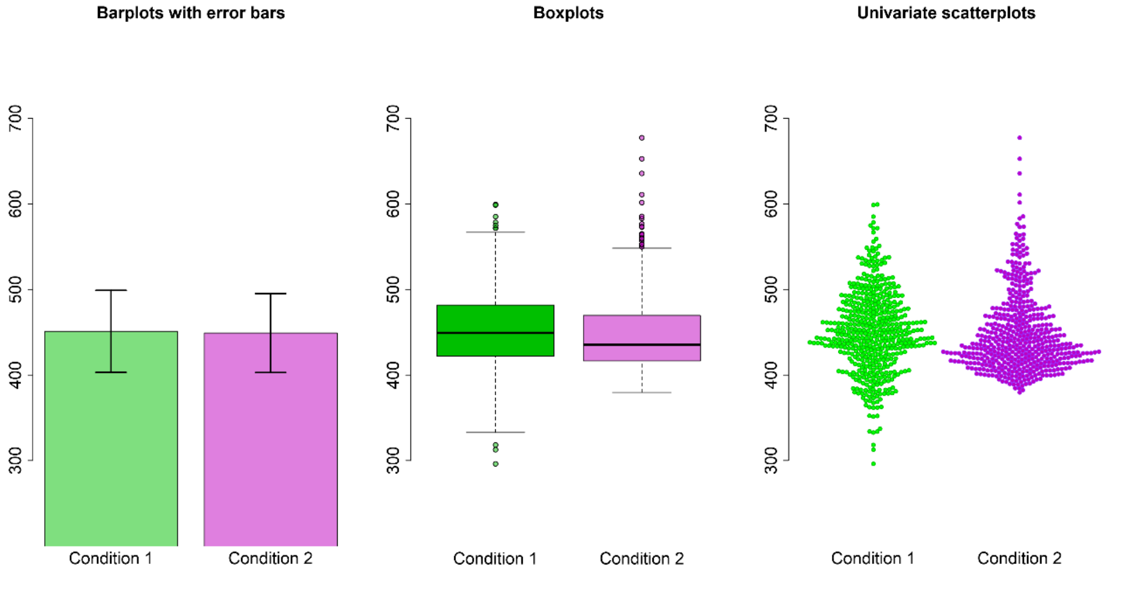

Figure 1. Three visualizations of the same dataset (based on the figure used by the #barbarplots campaign; https://barbarplots.github.io/). The data were simulated such that Condition 1 and Condition 2 would have the same mean, standard deviation, and sample size, but Condition 1 follows a normal distribution whereas Condition 2 follows an exponentially modified Gaussian distribution. The bar plot with error bars showing ±1 standard deviation (left) makes the two conditions look the same. The boxplot (middle) is suggestive of difference between the distributions. Finally, univariate scatterplots (also called strip plots) showing each data point in each condition (with the data points arranged along the horizontal axis to avoid overlap, and to mimic the shape of a histogram or violin plot, using the {beeswarm} package [Eklund, 2016] in the R statistical computing environment [R Core Team, 2016]) clearly shows the different shapes of the distributions.

There are many other situations in which two datasets may differ in important ways that are not revealed in a data summary that only shows a measure of central tendency and a measure of variance or precision. For instance, distributions with very different standard errors might be this way because they have different variances, or because one has a much bigger sample size, as shown in Figure 2. Two conditions may both have the same distribution, but a non-normal one (e.g., two datasets might both follow skewed distributions like that shown for Condition 2 in the univariate scatterplots on the right-hand side of Figure 1), in which case the mean may not be a very accurate summary of either condition's data. Skewed distributions like these are common for types of data that have a natural lower or upper bound, such as syllable durations or reaction times, neither of which can be less than zero. Because of the abovementioned limitations of plots showing simple summary statistics, recent authors have advocated the use of visualizations which show the full distribution of data (e.g., Rousselet et al., 2016; Weissgerber et al., 2015).

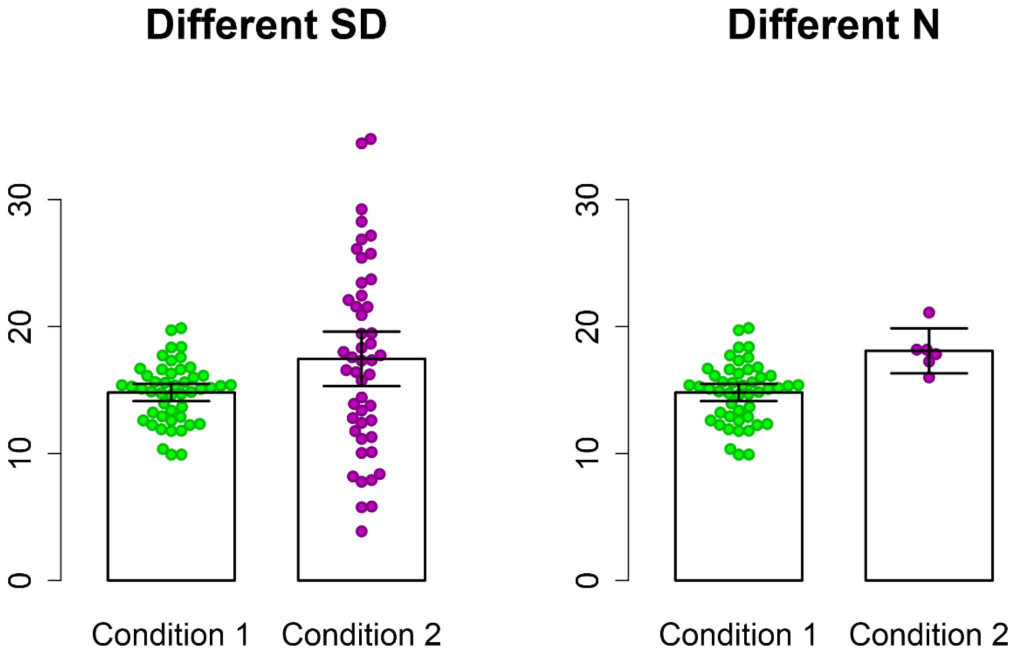

Figure 2. Illustration of two different sources of differences in standard error bars. In each panel, Condition 2 on the right has a wider error bar (indicating a 95% confidence interval of the mean, calculated based on standard error times the critical t*-value). The reason for the wider error bars, however, is different. In the left-hand panel, this difference is due to a larger standard deviation for Condition 2—the data were simulated from a normal distribution such that Condition 2 had the same sample size as Condition 1 but a larger standard deviation. In the right-hand panel, on the other hand, both conditions have the same standard deviation but Condition 2 has fewer observations—the data were again simulated from a normal distribution, this time to have the same standard deviation as Condition 1 but a smaller sample size. If the plot only showed error bars, it would not show which factor, standard deviation or sample size, is making one condition's error bar wider than the other's.*

These are just some of the reasons why plotting raw data, or at least fully exploring the distribution of data, is important. Thus, recent advice such as that by Rousselet and colleagues (2016) and Weissgerber and colleagues (2015), who advise (among other things) visualizing more complete distributions of data rather than just summary statistics, is not to be taken lightly. The main argument of this paper, however, is that in some situations this advice is impossible to follow; showing all the relevant information about a dataset at once is not actually possible for many research designs common in phonetics and psycholinguistics. In addition to the challenges acknowledged by these authors (for instance, that designs with many conditions to compare are difficult to show in a single visualization), there is also a fundamental challenge. In repeated-measures experiments, the structure of a dataset is more than just the raw values; the connections between data points in different conditions is just as important. Below we will illustrate why it is not possible to show all of these connections at once, and offer some suggestions for strategies to show as much of the key information as possible.

2.2 Accounting for paired data

Many research designs in phonetics and psycholinguistics use repeated measures, typically by participants—in other words, one participant contributes data to each condition. For example, in an acoustic experiment measuring voice onset time (VOT) in fast and slow speech, a given participant contributes at least two data points, one from a word produced in fast speech and one from a word produced in slow speech. (A given word also contributes multiple data points, as will be discussed in section 2.3.) In such a case, merely showing a scatterplot or histogram of the individual participants' data points for each condition (fast and slow) is insufficient; the pattern of within-participant changes must also be represented. Consider, for instance, data from two fake experiments shown in Figure 3, where the top row of figures shows one experiment and the bottom row another. These experiments have the exact same data points, and indeed the bar plots of the means and error bars showing the 95% confidence intervals of the mean of each condition (in black) are the same, suggesting that the results are identical across the two experiments. In many research reports in phonetics and psycholinguistics, all that is shown is a visualization of central tendency (e.g., the mean) and dispersion or precision (e.g., standard deviation or standard error). However, the pairing between the individual data points in fast and slow speech is different across the two experiments. In Experiment 1 (top row), almost every participant showed a slightly longer VOT in slow speech than in fast speech. In Experiment 2 (bottom row), the differences between fast and slow speech vary widely across participants. Accordingly, a paired *t*-test for the difference of means in Experiment 1 is highly significant ($t(19) = -3.34$, $p = 0.003$), whereas in Experiment 2 it is negligible and non-significant ($t(19) = -0.60$, $p = 0.554$).

In short, plots of condition means and confidence intervals hide very important differences. The only way to accurately represent the repeated-measures pattern is to show the paired observations, as done here with the lines connecting the data points, or to show the participant-wise differences, as shown on the right-hand side of each row in red dots. This problem has been known for a long time; see, for instance, Loftus & Masson (1994) and Weissgerber, Milic, Winham, and Garovic (2015) for similar demonstrations of how misleading confidence intervals of condition means can be for paired data. Alternative types of intervals for the display of within-participant data have been proposed (see Baguley, 2012, for review). Rousselet et al. (2016) give a useful overview of the types of patterns that can be revealed through informative plots of paired data. The important lesson is that when there are repeated measures, a plot of two or more conditions and their associated standard errors or confidence intervals is uninformative with respect to whether or not those conditions significantly differ. While people often attempt to give such plots heuristic interpretations (e.g., "these conditions don't differ because their error bars overlap", etc.), such conclusions are not valid, as Figure 3 demonstrates.

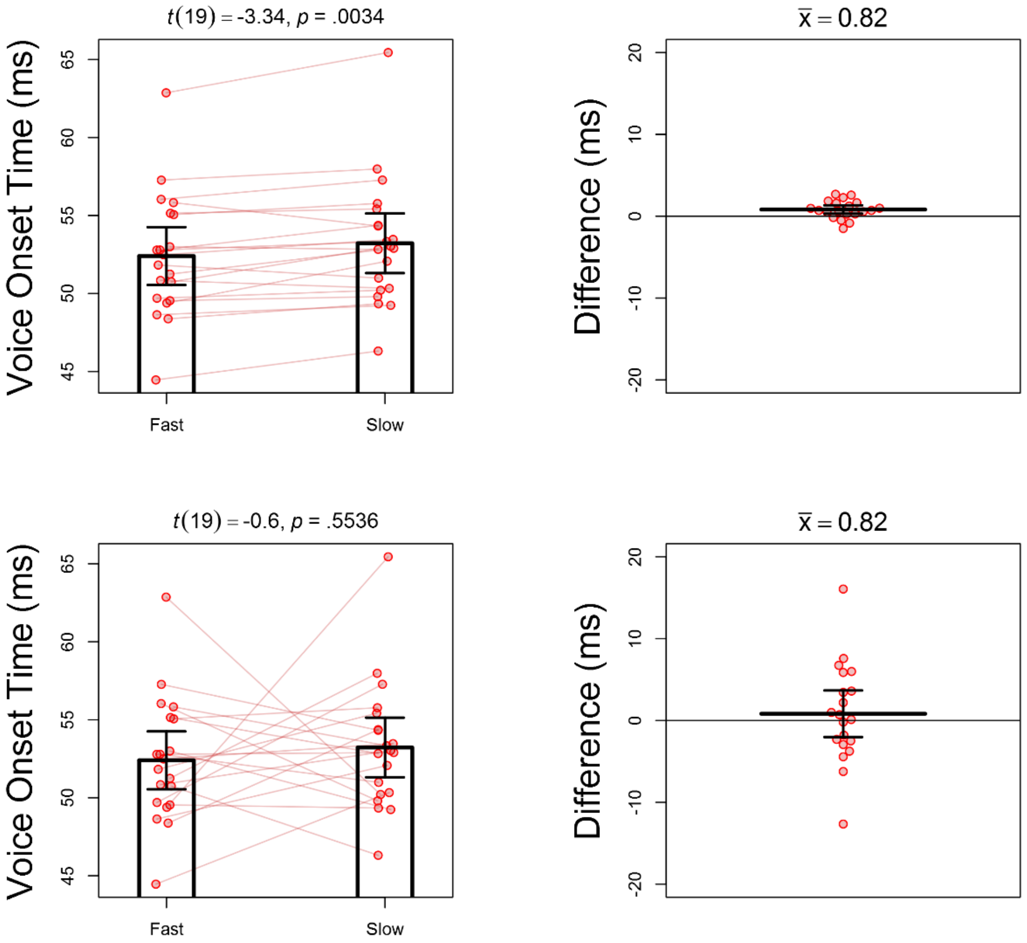

Figure 3. The same two series of data points, either from an experiment where the within-participant pairing of data points was consistently positive (top two panels) or where the within-participant pairing was extremely variable (bottom two panels). Error bars represent 95% confidence intervals based on the standard error of the mean and the t*-statistic. Points are jittered on the horizontal axis to reduce overlap. See text for details.*

2.3 Accounting for crossed random effects

Visualizations that plot differences or paired observations are easily implemented (albeit not widely adopted in phonetic sciences). The above recommendations are particularly useful for experiment designs in which participants are the only variable measured repeatedly. In many phonetics and psycholinguistics experimental paradigms, however, there is more than one repeated-measures factor. For instance, stimuli may, just like participants, contribute data points to each condition in an experiment. Consider again our example with speech rate and VOT. In such an experiment, typically speaking, not only will data be collected from multiple participants, but also from multiple words. For instance, in a completely repeated design, one person will produce words like “cat” and “cab” in two conditions each, once each in the fast-speech condition and once each in the slow-speech condition. In a Latin square design, half of the participants will produce “cat” in the fast-speech condition and “cab” in the slow-speech condition, and the other half of the participants vice versa. Either way, each word in the experiment will be produced multiple times, and thus will contribute multiple data points to the results. Therefore, in addition to repeated measures for participants, there are also repeated measures for these words (hereafter referred to as "items").

The use of multiple items is crucial for experiments interested in making generalizations about language. For example, if a researcher wants to demonstrate that speech rate affects VOT systematically in a language, rather than just idiosyncratically affecting the VOT of one or two words, it is necessary to conduct an experiment with a large sample of words, to allow for making statistical inferences about the population of words in a language. The use of different items is also often necessary for experimental design reasons; for example, phonetic and psycholinguistic data are often noisy, necessitating the collection of many trials of data to obtain an acceptable signal-to-noise ratio, but repeating the same item over and over again would lead to familiarization, fatigue, or other repetition effects. As a result, the best option is to include many different items. For reasons like these, experiments including both multiple participants and multiple items have become the norm in phonetics and other linguistic subfields, and thus making statistical inferences about items as well as participants is necessary.

For example, for another simulated dataset from a fully repeated design like this, Figure 4 demonstrates two ways to present these data: one summarizing points by participants, and one summarizing by items.

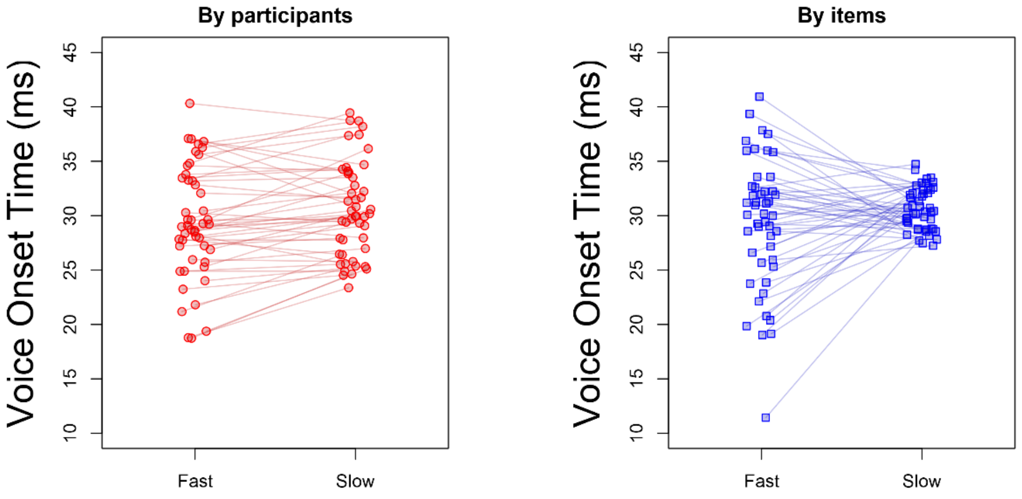

Figure 4. Plot of paired data from a simulated dataset, aggregated either by participants (left panel) or by items/stimuli (right panel).

Figure 4 illustrates data from a design like this, where VOT is measured in fast and slow speech and the design is fully repeated within participants and within items. Such designs pose a special problem for the visualization of paired data. It is no longer possible to plot a single data point or a single pair of data points for each participant, since the participant has contributed many data points. Instead, it is generally necessary to aggregate over all that participant's observations for each condition, as shown in the left-hand side of Figure 4. For instance, one participant's context effect in the experiment described above would be their mean VOT for fast-speech productions minus their mean VOT for slow-speech productions. This analysis discards a substantial amount of data, however, because it averages over variation in stimuli. For example, in the dataset illustrated in Figure 4, aggregating over participants obscures the fact that VOT varies less across items in slow speech than it does in fast speech; that is only visible on the right-hand side of the figure, where data are aggregated across participants rather than across items.

In short, there are two ways to look at the data: averaged across items for each participant, or averaged across participants for each item. The former ignores the variation across items and the latter ignores the variation across participants. These analyses also ignore the fact that within participants and items there may be different distributions. For example, two participants could have the same mean difference between conditions, but different standard deviations for each condition. Thus, plots like these, which condense a given participant's or a given item's multiple observations into one data point, discard potentially valuable information.

A related issue is that for any given participant or any given item, aggregation may mask or misrepresent important aspects of the data. The mean may not necessarily be the best representation of a given participant's or item's data points; while Figure 1 demonstrates that mean and standard deviation may not provide an accurate summary of a whole dataset, the same issue applies to the set of observations for a single participant or a single item. A given participant or a given item might show differences between conditions that are not captured by the mean. In some cases, each participant's or item's data might be better summarized by a different location parameter such as a trimmed mean or a median (which is just an extreme version of a trimmed mean; Rousselet, 2017).

Finally, test statistics calculated on datasets aggregated by participants or by stimuli often yield inappropriate inferences, a problem that has been long known (e.g., Clark, 1973; for recent treatments see Baayen, Davison, and Bates, 2008, and Judd, Westfall, and Kenny, 2012). While this issue is prominent in phonetic and psychological research on language processing, it is also relevant for many other aspects of cognitive and social psychology in which experiments typically use within-stimulus designs (see Judd et al., 2012, for examples). Indeed, the problem posed by this sort of design is one reason that LMEMs have become so popular in recent years. Our plotting standards, however, have yet to match these new models.

Figure 5 illustrates one way that aggregating by participants or stimuli can yield inappropriate statistical inferences. This plot shows the same speech rate dataset illustrated in Figure 4. The left-hand univariate scatterplot shows the pairwise difference for each participant, along with the 95% confidence interval (based on the standard error of the mean) of the mean of these differences. The 95% confidence interval does not include zero, suggesting that the difference between conditions is statistically significant at the α=.05 level. However, the right-hand scatterplot shows the pairwise difference of each item, which is substantially more variable. There the 95% interval includes zero, suggesting that the difference between conditions is not statistically significant. In fact, when analyzing the results with an LMEM with a fixed effect of speech rate condition and maximal random effects for both participants and items (Baayen et al., 2008; Barr et al., 2013; Judd et al., 2012),[1] the difference between conditions is not significant (*b* = 15.44, 95% percentile bootstrap CI = [-11.68, 44.20], *t* = -1.10). Thus, showing a plot of only the by-participant paired observations or pairwise differences would give a misleading visualization of the data. Likewise, because of this issue, solutions that have been suggested to improve the presentation of data (i.e., plotting individual data points or individual participant patterns; Rousselet et al., 2016; Weissgerber et al., 2015) are not directly applicable to many experimental designs.

[1] Model formula in R's {lme4} syntax: `VOT ~ Condition + (1+Condition|Participant) + (1+Condition|Item)`.

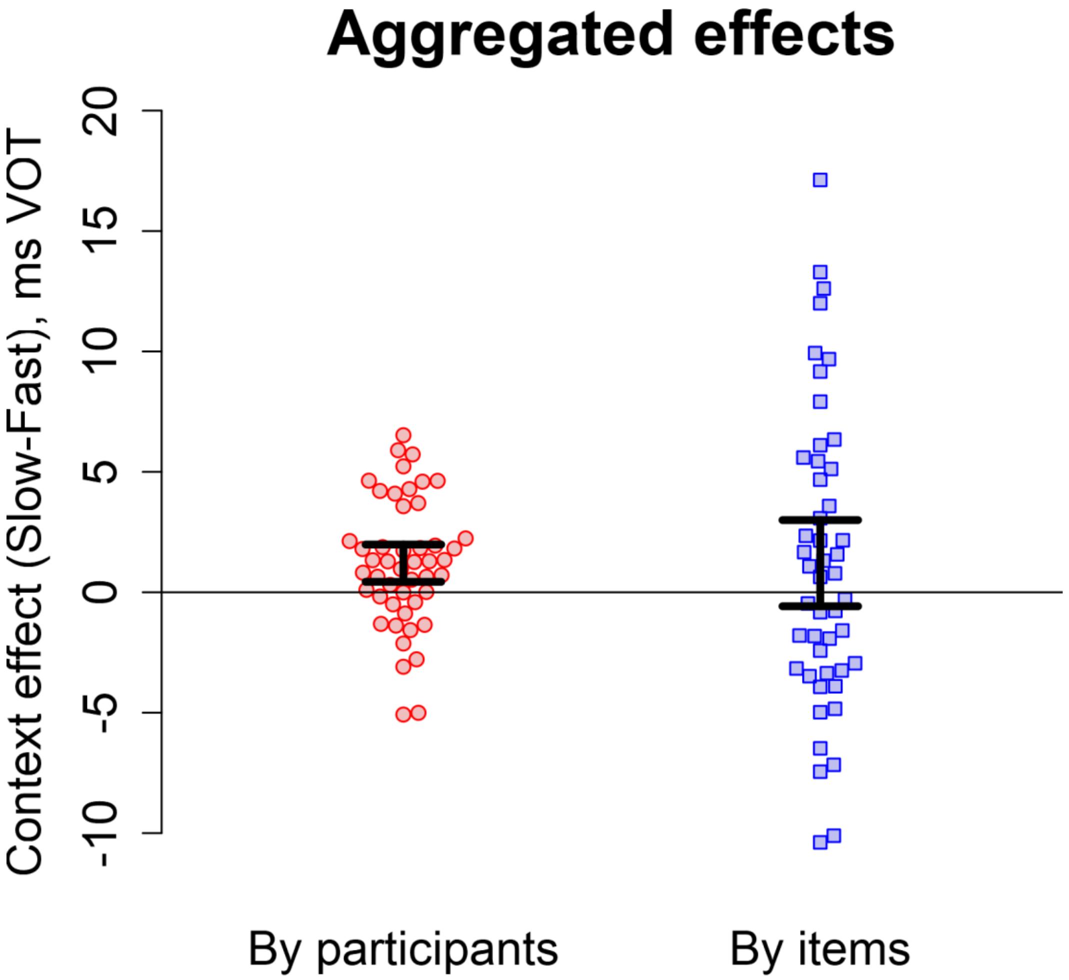

Figure 5. By-participant and by-item aggregated differences (mean VOT in slow context minus mean VOT in fast context), along with 95% confidence intervals (error bars) of the mean of the differences (calculated based on the standard error and critical t-value).

2.4 Alternatives that are also problematic

As shown above, plots that aggregate over items or over participants do not show all of the potentially important information in a dataset. Thus, while visualization in phonetics would be greatly improved by adopting practices recommended by the likes of Rousselet and colleagues (2016) and Weissgerber and colleagues (2015), datasets with repeated measures for both participants and items present additional problems which are not surmountable by any of those practices. Here we briefly consider whether any other techniques can easily overcome these problems.

One tempting option is to make unique histograms (or kernel density/violin plots, or scatter plots) for every level of our random effects. If aggregating over items or participants is problematic, why not forego aggregating? For example, a plot could show the individual data points for each participant separately, as well as another set of individual data points for each item separately; these are sometimes called "small multiples" plots. An example of this is shown in Figure 6. This gives a rich picture of the data and allows the researcher to see whether any particular participant or item shows a difference between conditions in some aspect of the data distribution other than the mean. In a plot like this, however, it is difficult to get a clear sense of what one is trying to say with this data—in other words, a visualization like this makes it difficult to summarize the data at a glance. For instance, the variance in item effects is difficult to see when data are spread across so many subplots. Plots like these may be useful during an initial exploration phase, but would most likely not be appropriate for dissemination in a final report. Furthermore, they run up against the last problem described above: the by-participant plots fail to show the repeated nature of the item observations, and the by-item plots fail to show the repeated nature of the participant observations.

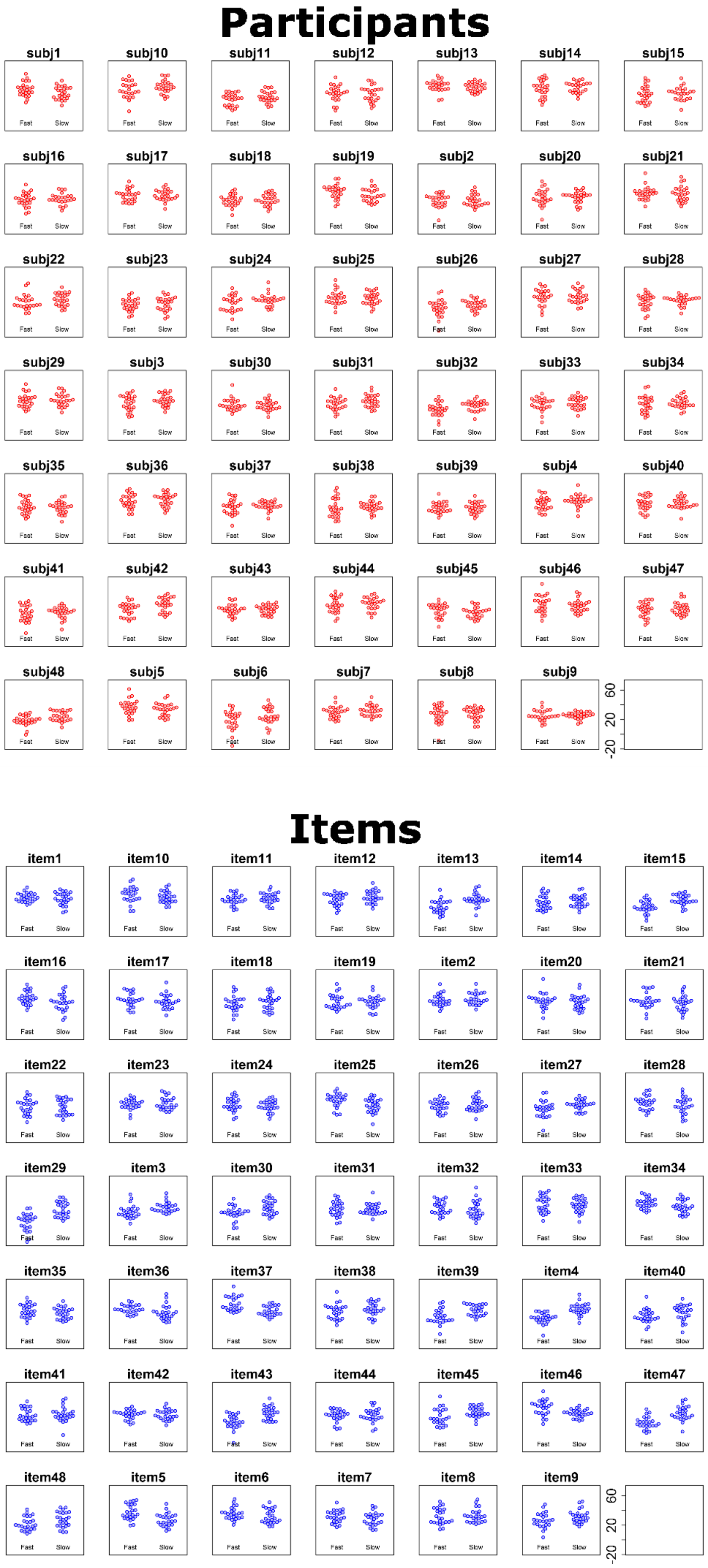
Participants
subj1
subj10
subj11
subj12
subj13
subj14
subj15
subj16
subj17
subj18
subj19
subj2
subj20
subj21
subj22
subj23
subj24
subj25
subj26
subj27
subj28
subj29
subj3
subj30
subj31
subj32
subj33
subj34
subj35
subj36
subj37
subj38
subj39
subj4
subj40
subj41
subj42
subj43
subj44
subj45
subj46
subj47
subj48
subj5
subj6
subj7
subj8
subj9
Fast
Slow
-20
20
60
Items
item1
item10
item11
item12
item13
item14
item15
item16
item17
item18
item19
item2
item20
item21
item22
item23
item24
item25
item26
item27
item28
item29
item3
item30
item31
item32
item33
item34
item35
item36
item37
item38
item39
item4
item40
item41
item42
item43
item44
item45
item46
item47
item48
item5
item6
item7
item8
item9
Fast
Slow
-20
20
60

Figure 6. Univariate scatterplots of VOT (in milliseconds) for each participant and item, by condition. The last subplot in each section provides a scale on the vertical axis.

While plotting without any aggregation can be confusing and obscure results, as shown above, focusing too much on model outputs can also be problematic. For example, an option that may seem tempting is to plot random coefficients from an LMEM, rather than plotting participant-wise or stimulus-wise means. An LMEM returns not just estimates of fixed effects, but also *best linear unbiased predictors* (BLUPs) representing how much that estimate differs, for each participant or each item, from the overall fixed-effect estimate (Baayen, 2008; Blouin & Rioppele, 2005). Adding the BLUPs to the fixed effect yields different coefficients for each participant or each item; importantly, these BLUPs take into account both participant and item variability at the same time, since they come from a mixed-effects model. In this sense, they might be preferable to simply plotting participant-wise or item-wise means, which aggregate over variability in items or participants. By using BLUPs from a mixed model, we could plot each participant's effect of speech rate in a way that takes item variability into account, and each item's effect of speech rate in a way that takes participant variability into account.

This option is still problematic, however, since it is showing model parameters rather than data. This is the same concern we have raised with other methods: a plot of participant-wise means is essentially a plot of coefficients from a simple model in which participants are treated as a fixed effect and thus each have their own intercept. Plotting summary parameters rather than data always carries limitations, regardless of whether these summary parameters are means or BLUPs. Summary parameters are a simplification (albeit a sometimes necessary one) of the data, and they are highly susceptible to the details of the model from which they were taken. This is particularly the case for mixed-effects models, where there are many different ways the model can be specified, particularly when it comes to the model's random effects structure. Different model specifications may result in very different coefficients.

For instance, Figure 7 shows the model coefficients from two different models with different specifications. Imagine that the dependent variable is VOT and the independent variable is how much time the participant has to speak: the longer time they speak, the slower the speech rate can be.[2] () The dots in each plot show each participant's speech rate effect, where a positive effect means that VOTs got longer as the participant spoke more slowly. The left panel shows coefficients from a model in which parameters for correlations between random intercepts and random slopes have been suppressed. Suppressing these correlation parameters is a common practice used to help LMEMs converge more quickly and avoid convergence failures (Kliegl, 2014). In this model, all but two participants show positive effects of speaking time—in other words, the more slowly they speak, the longer their VOT becomes. This pattern might lead the researcher to believe that there is a reliable trend in this direction. The right panel, however, shows coefficients from a full model which includes random correlation parameters. In this model the mode of the distribution of participant-wise coefficients is approximately zero, which might lead the researcher to believe that there is little evidence for an effect in this direction. For these reasons, a plot like this is not ideal for visualizing and exploring a dataset (as opposed to visualizing and exploring a model). While all plots of anything other than raw data are similarly susceptible to modeling choices (e.g., when plotting participant-wise means, the plot will be influenced by the researcher's choice to model the data using by-participant means rather than by-participant medians), the number of choice points in constructing a mixed-effects model is substantially larger, compounding this concern.

[2] This is just a simplification to keep this example consistent with the other examples used in this paper; in reality the data are not from a VOT experiment; they are from an unrelated experiment examining whether the difference in sonority between two consonants in a branching onset influences the acceptability of a pseudoword (Berkson & Flego, 2017).

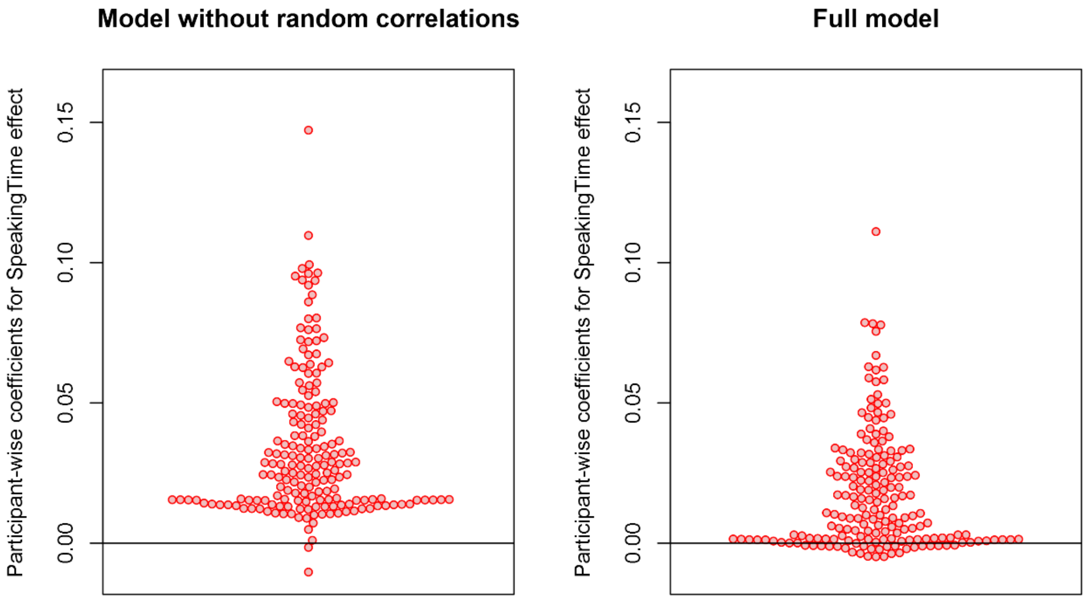

Figure 7. By-participant coefficients (fixed effect plus best linear unbiased predictor) from two different models of the same dataset. Each point represents one participant, and the width of the scatterplot is a function of kernel density (as in a violin plot).

An additional problem with using BLUPs is that they undergo shrinkage (Blouin & Rioppele, 2005); that is to say, in a mixed-effects model, the effects for each participant and item are not considered to be independent, but are considered to be perturbations around a common fixed effect. Therefore, the coefficients estimated for each participant and item in a mixed model will be closer to the overall mean coefficient (the fixed effect) than they are in reality. This is shown in Figure 8, from the fake VOT data used in the previous examples; the mixed-effect model "moves" each participant's estimated effect closer to the fixed effect.

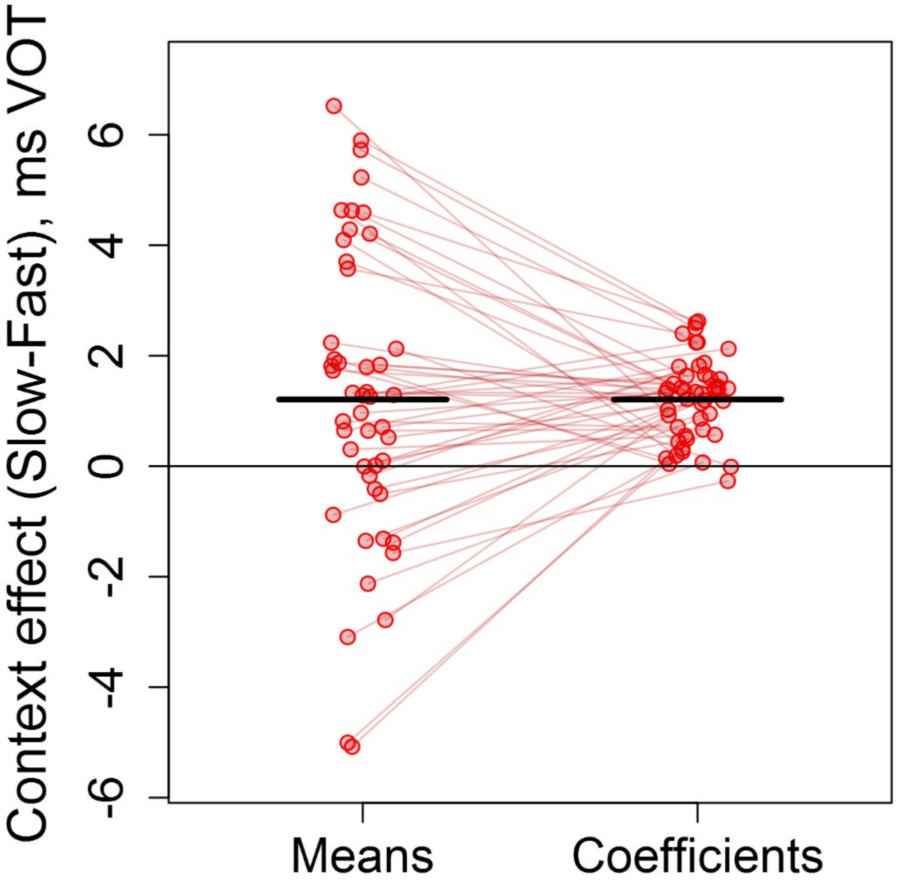

Figure 8. Difference between actual aggregate values and coefficients estimated from a mixed model. Each pair of dots shows one participant's mean difference between fast and slow speech (left side) and the same participant's difference estimated from a mixed-effects model (the fixed effect plus that participant's BLUP). The black lines represent the mean difference (left) and the fixed-effect estimate of the difference in a mixed-effects model (right). The model coefficients for the participants are more tightly clustered around the mean difference than the actual observed mean differences are.

In short, neither of the alternatives considered above (small multiples plots or plotting BLUPs rather than by-participant and by-item means) can satisfactorily resolve the problems we have raised. To our knowledge, there is no existing plotting technique that can; for the kind of repeated-measures data commonly used in phonetics and psycholinguistics, any plot that one can make will necessarily obscure at least one aspect of the data. Whereas contemporary phonetics and psycholinguistics research is increasingly making use of statistical models that can handle all sources of variability at once, we do not have any plotting technique that can similarly represent all the relevant variability in one plot. Thus, there is no single solution to the problems presented above; the best a researcher can do is to use multiple plots to show different aspects of the data. It follows that there is not one ideal way to present data, but that different plots will be ideal for different purposes. In the final sections of this paper we go over what these different purposes are, and some recommendations of how to approach data visualization in future projects.

3 What is the purpose of a visualization?

In the previous section we reiterated that no single plot can solve all the problems we have outlined, and that different plots must thus be used to accomplish different purposes. Here we review what some of these different purposes are. It will be instructive to illustrate these purposes using some more complex research designs. The examples in the preceding section were all based on one of the simplest possible designs in an experiment with crossed random effects: two conditions, to which each participant and each item contributes observations. In such a design there is only one comparison of interest. When plotting a design with more potential comparisons, the problem is compounded: not only can no comparison be fully illustrated (for the reasons described above), but also the issue of which comparison should be shown is raised.

Consider a design with three conditions. For instance, imagine in our fake VOT experiment we were also interested in comparing VOT for stops at different places of articulation. Figure 9 shows data simulated from such a design. Each point represents one participant's average VOT for words at that place of articulation. The left and right halves of Figure 9 show two different ways to arrange the data for visualization. The plot on the left shows that most participants had longer VOTs for coronal than labial stops, and almost every participant had longer VOTs for velars. However, a direct comparison between labials and velars is not shown. In this particular experiment we can easily infer that just about every participant has longer VOTs for velars than for labials, since they have longer VOTs for coronals than labials and longer for velars than coronals. But in another experiment where the data pattern is less robust, this would not be so easy to visualize, and then the choice of the order in which to arrange the conditions on the plot would have substantial consequence. The right-hand side of Figure 9 shows a consequence of different arrangement; here it is not easy to see how many participants had longer VOT in coronal than in labial stops. The only way to show all three possible comparisons would be to plot the pairwise differences (as in Figure 5) rather than the values for each condition itself (as in Figures 4 and 9).

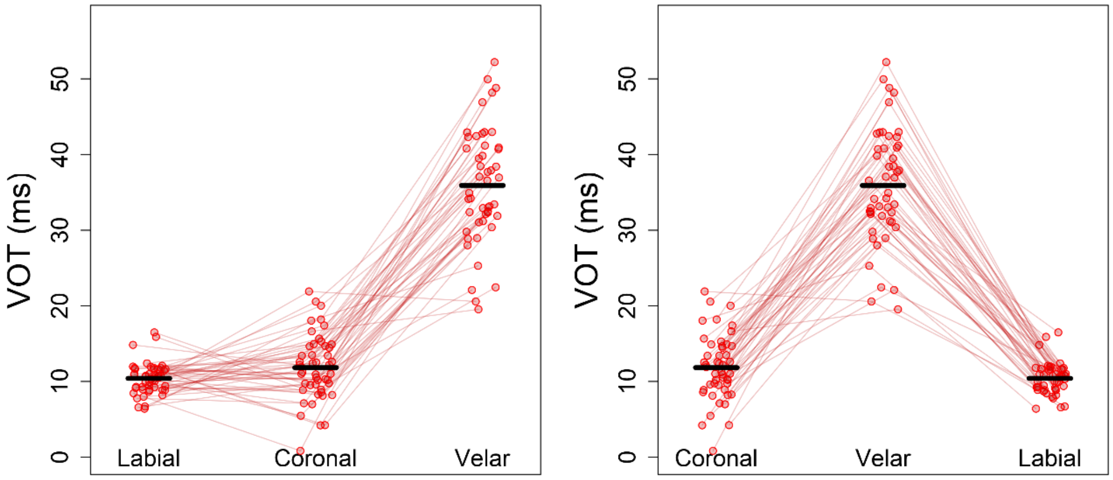

Figure 9. VOT data for stops at different places of articulation. Each red circle represents the mean for one participant, and black lines represent the mean VOT for the condition across all participants. The two panels illustrate two different arrangements of the same dataset.

When making such decisions about what a plot should illustrate, it is necessary to consider what the purpose of the plot is. Thus far this paper has mainly concerned itself with plots functioning as aids for statistical inference: we have focused on whether or not a given visualization appropriately represents whether two conditions differ in terms of their distribution, whether a difference is consistent across participants/items, and whether the conclusions suggested by a plot are congruent with the conclusions suggested by inferential statistics. Indeed, an ideal approach to statistical inference may be to forego making dichotomous decisions based on *p*-values (a practice based on Neyman-Pearson hypothesis testing [Neyman, 1957], but substantially altered from what they originally intended; see Gigerenzer, 2004, for details), and to instead judge how robust a pattern is based on how it appears in visualizations like these (e.g., Rousselet et al., 2016); this has been jokingly termed the "inter-ocular trauma test" (examining whether a plot "hits you between the eyes", i.e., whether a pattern is visually obvious). However, a plot can also serve other functions, such as to help explore data (e.g., for identifying outliers or patterns of potential interest) and to help communicate a point (e.g., to show some audience an important pattern in the dataset). Let us consider a dataset that illustrates how different visualizations can serve some of these different functions.

Imagine that we conducted another experiment like the one described above to examine VOT in slow and fast speech, but also tested this effect at three different places of articulation: labial, coronal, and alveolar. Figure 10 illustrates three different ways we could visualize data on VOT in these six contexts.[3] The top section of Figure 10 shows the sort of plot most commonly used to illustrate such a 2×3 factorial design. The middle section shows the by-participant and by-item differences, as in Figure 5. The bottom section shows a version of the first plot (from the top section of the figure) with individual participant data added, like Figures 3 and 9.

[3] The actual data are taken from an unrelated psycholinguistic experiment on reading (Politzer-Ahles & Husband, 2018); the condition labels have been changed for the purpose of this example.

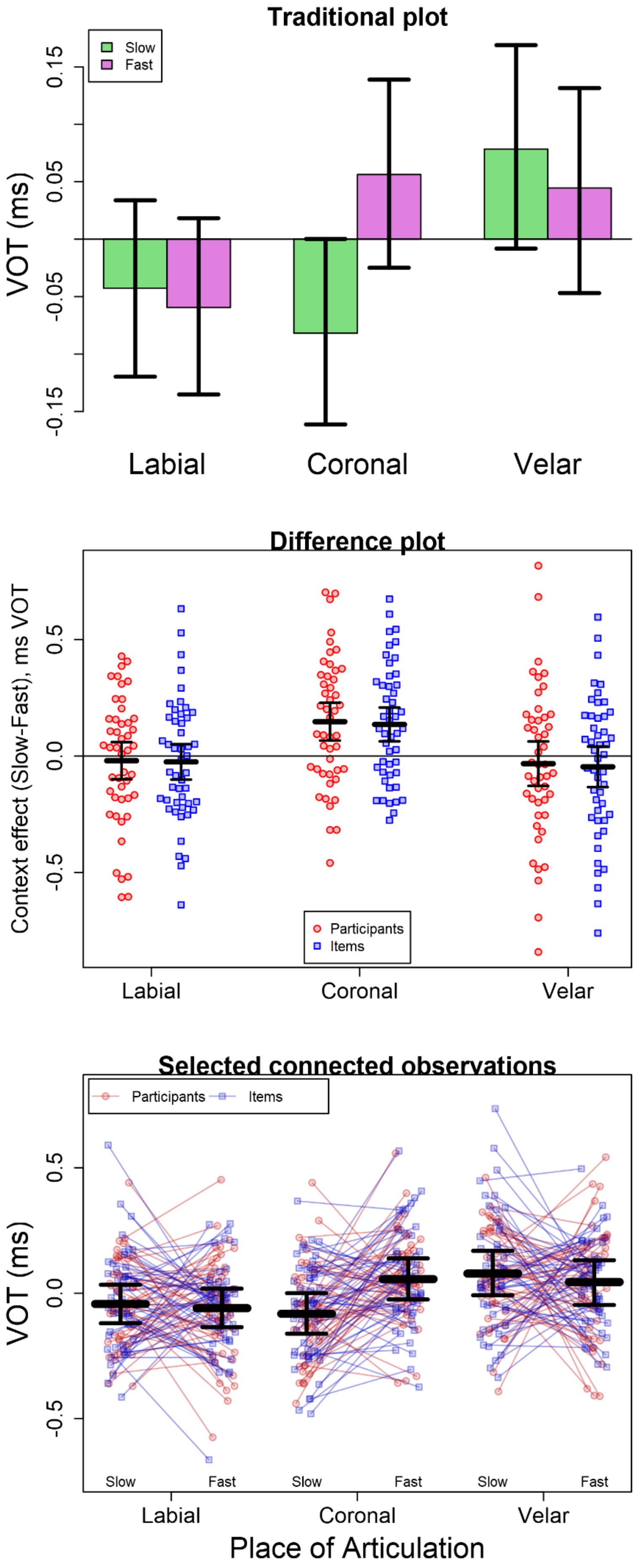
Traditional plot
Slow
Fast
VOT (ms)
0.15
0.05
-0.05
-0.15
Labial
Coronal
Velar
Difference plot
Context effect (Slow-Fast), ms VOT
0.5
0.0
-0.5
Participants
Items
Labial
Coronal
Velar
Selected connected observations
Participants
Items
VOT (ms)
0.5
0.0
-0.5
Slow
Fast
Slow
Fast
Slow
Fast
Labial
Coronal
Velar
Place of Articulation

Figure 10. Three different visualizations of the same data. See text for details.

The first plot suggests that there is a significant difference between slow and fast speech for coronals, but not for the other places of articulation.[4] Furthermore, it reveals a larger pattern: the first three conditions pattern together, and the second three conditions pattern together, and these two clusters of conditions are different from one another. Or to put it another way: coronals in slow speech pattern like labials, (regardless of speech speed), and coronals in fast speech pattern like velars (regardless of speech speed). However, this plot has all the limitations mentioned in the discussion of Figures 1 and 2: it does not show information about individual participants or the shape of any distributions in the data. This is an effective plot for the function of communicating a pattern, but not for the function of aiding statistical inference.

The second plot shows by-participant and by-item pairwise differences (slow speech minus fast speech) for each place of articulation. In this case the error bars represent standard confidence intervals—for a given difference, the difference is significantly different from any value outside the confidence interval. Therefore, this plot, like the plot above it, suggests that there is a significant effect of speech rate in coronals, but not the other two places of articulation. Unlike the first plot, it also shows more detail about the pattern of these differences; rather than just making a dichotomous decision about whether the difference is statistically significant, the viewer can judge how strong the pattern is or how robust across participants and items by, for example, seeing how many participants and items show an effect in the direction suggested by the overall mean. On the other hand, this plot also hides some things that the plot above it does not—most importantly, the generalization that coronal-slow patterns with labial while coronal-fast patterns with velar. Furthermore, as described above with respect to Figure 9, this plot only shows three pairwise comparisons, out of a possible 15 (this design has six conditions, and in any K conditions there are $\frac{K(K-1)}{2}$ possible pairwise comparisons, not to mention main effect or interaction comparisons where relevant). The comparisons to be shown may have been chosen on a logical basis—for example, maybe they were the only comparisons that were important for the research question—but this is done at the cost of obscuring other aspects of the data. Thus, while this plot effectively serves the function of aiding statistical inference (albeit only for the specific comparisons the creator has decided to illustrate), it does not effectively communicate the pattern shown in the first plot. The plot is also effective for some kinds of data exploration (e.g., seeing whether some participants or items stand out), but not for others (e.g., making comparisons other than the one that the creator of the visualization intended). The plot's function for aiding statistical inference is also limited by the drawbacks discussed above—that it requires aggregating over items or over participants.

The bottom plot shows one way to compromise between these inferential and communicative functions. The solid black lines in the plot show the same information as the bars and error bars in the top plot; thus, this plot can also show the same general pattern as the first one. Points for individual participants and items are overlain, to help illustrate the extent to which these differences are or are not robust across participants and items. For example, the context effect is significant for coronals and, accordingly, most of the lines from slow to fast are going in one direction (up), but nevertheless there are also a good number of lines going the opposite direction, consistent with the middle plot which shows that there are still many participants and items with effects in the "wrong" direction. Thus, this plot combines information from both of the other plots. It is not without its tradeoffs, though. To accommodate the full range of individual participant and item means, the y-axis is substantially lengthened, compressing the means and error bars together; this makes the general pattern less obvious than it was in the top plot, detracting from the plot's effectiveness for communication. At the same time, counting the number of upward-sloping versus downward-sloping lines in this plot at a glance is much more difficult than seeing the number of points above and below zero in the

[4] The error bars in the plot represent difference-adjusted intervals, based on a linear mixed-effects model, that help illustrate which pairwise comparisons are significant or not (Politzer-Ahles, 2017). For any given pair of conditions, if one condition's mean is outside of the other condition's error bars and vice versa, they are likely to be significantly different.

middle plot; this makes the plot less effective for easy statistical inference. Like in the second plot, many possible pairwise comparisons are not shown. Finally, the general complexity of the plot makes it harder to take in quickly; a plot like this is probably more appropriate for the researcher's own exploration and for papers in which the reader has time to pore over the plot, and less appropriate for presentations in which an audience needs to understand the results and conclusions in a limited amount of time.

The examples above illustrate that the same set of data may need to be graphed in different ways depending on the purpose. The issues raised in section 2—that not all aspects of data with crossed random effects can be shown in a plot—are not relevant if the primary purpose of a plot is to convey a general data pattern rather than to aid statistical inference. Likewise, if a plot is made to aid statistical inference and alleviate the concerns described above, this may come at the cost of not clearly showing the general data pattern across many conditions. In what follows, we will continue to focus on a simple case of two conditions and one comparison of interest. For most real-life datasets, this means that the dataset may need to be broken down into one or more discrete comparisons of interest (such as in the bottom two plots of Figure 10) if one wants to make statistically informative visualizations.

4 Recommendations

Researchers often make plots after running their statistics. Recently, there has been more of a push to make plots first, in order to explore the data, diagnose models, and be sure the correct test is being used for the type of data available. However, this can lead to very complicated visualizations, and when it comes to communicating these data for an outside audience, one often needs to present a simpler summary. It is necessary, therefore, to have multiple stages of visualization to give the most complete picture of the data, and analysis, as possible. While the details of what plots are built will vary depending on the research design and aims, below we present what we consider an ideal general approach to visualizations to use, particularly for research that is at least partially exploratory. The general steps are as follows: (1) examine the data distribution within each unit (i.e., each participant, each item); (2) aggregate over items or participants to plot differences or paired data points for each comparison of interest; (3) make simpler plots for communicative purposes; and (4) provide detailed plots, or raw data, for posterity.[5]

4.1 Examining the data distribution within each participant and within each item

Figure 6 is an example of this sort of plotting following the first step in the guidelines above. Examining the data at the individual level is useful for understanding the distributions that each participant's or item's data follows—in other words, for seeing features of the data that will be lost once we move into the realms of looking at summary statistics like by-participant means. In this way researchers may discover unanticipated but potentially important aspects of the data distributions. Even if such differences are not the primary focus of the research question, they may be relevant for other research questions, and no one would become aware of them if only data summaries were ever shown. For instance, Staub and colleagues (2010, 2013) did several analyses testing whether the skew of reading times in eye-tracking experiments (i.e., the extent

[5] It is important to note that these steps do not necessarily need to all be followed for every research project. Particularly in the case of pre-registered, confirmatory research, where just one pre-determined aspect of the data is of interest, there may not be much need to explore data distributions, for example. If the research hypothesis is specifically about a difference in one parameter (such as the mean), there is no need to go fishing for other differences in data distributions. When interesting patterns are revealed unexpectedly during data exploration, these should be clearly presented as exploratory rather than confirmatory (Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Nevertheless, these sorts of detailed plots or raw data always should be available for other exploratory analysis in the future; for example, even if a researcher conducts a pre-registered confirmatory experiment interested only in condition means, years later they might become interested in whether or not there are other differences in the distribution (the research by Staub and colleagues [2010, 2013] discussed here is such a case).

to which there is a long right tail of slow reading times) is modulated independently of the mean reading times when various lexical factors are manipulated. Crucially, one of the motivating factors for those analyses was that previous reports had noted the skewed distribution of reading times (as mentioned by Staub et al., 2010); this illustrates the importance of making sure this information is available.

Even if the data do not show any potentially interesting differences in distribution between conditions, it is still useful to have a sense of whether or not each participant's or item's data shows a relatively normal distribution. While a normal distribution is not necessary for statistical models—as the general linear model only assumes normality in model residuals, not in the data themselves—it is still useful for interpreting model parameters or summary statistics like by-participant and by-item means. If data are severely non-normal, this may be a warning that parameters represented by summary statistics, like means and differences of means, are not accurate summaries. Likewise, if the data are seriously non-normal, then by-participant or by-item means might not be the best way to visualize how the effects vary across participants and items, since these means may not accurately depict any given participant's or item's distribution.

Plotting the data for each participant and for each item is also useful for other things, such as detecting outliers, if the researcher is not using another automatic outlier detection procedure. For instance, one outlier detection technique recommended by Baayen (2008:265-266) is to examine the raw data for each participant and each item and identify data points that visually stand out; this can be done with a plot showing every participant and item together as in Figure 6, or by plotting each participant and each item one at a time.

4.2 Paired data points or differences for comparisons of interest

As described above, a visualization like Figure 6 is not very useful for aiding statistical inference or understanding the gross pattern of differences in the dataset. For aiding statistical inference, we need a visualization that shows the paired differences for each participant and for each item. Depending on the point that the researcher wants to communicate, this might be done with a plot showing paired points for each condition, connected by a line, as in the top of Figure 11; or by a plot showing a single point for each pairwise difference, as in the middle of Figure 11; or by a two-dimensional graph, as in the bottom of Figure 11. As described in Sections 2 and 3, each of these plots has limitations – they require aggregation which obscures the underlying data distribution, and they privilege the display of one comparison while neglecting others (if there are other conditions in the experiment). However, they show as much information as possible, using the plotting techniques available. Figure 11 demonstrates three different ways our fake VOT data could be plotted—and supplemented with summary statistics from an LMEM—to facilitate statistical inferences and understanding of the important comparison in the dataset.

The "connected points" plot at the top shows each participant's and item's data in each condition. It is the most similar to a traditional bar plot (e.g., the thick lines representing the means show which condition has a higher mean VOT, and the error bars help show that the conditions do not significantly differ) and it also shows some distributional information clearly: it shows that items have less variance than participants in slow speech but not fast speech. However, it is difficult to visually estimate what proportion of the lines are upward-sloping and what proportion are downward-sloping; therefore, it is hard to make statistical inferences except by simply trusting the statistical model parameters (on which the error bars are based).

The "pairwise differences" plot in the middle shows each participant's and item's pairwise difference (slow speech minus fast speech). It clearly shows that a majority of participants and items have longer VOTs in slow speech than in fast speech, but that a substantial number also show effects in the opposite direction. The error bars also demonstrate well the difference between conditions. The result also fairly make clear that the effect is more variable across items than it is across participants. Unlike the "connected points" plot, however, it does not show that this is due to less item variance in the Slow condition specifically. It also requires slightly more explanation than the first—the reader has to keep in mind that the difference being shown is Slow minus Fast, not Fast minus Slow.

The “two-dimensional scatter”, based on Rousselet (2016), is probably the richest representation of these data. It shows both the proportion of participants and items with effects in each direction like the “pairwise difference” plot does (points above the diagonal are ones where slow speech has a longer VOT than fast speech, and points below vice versa; we can see that items spread out farther from the diagonal than participants do, indicating that they show more variance in their effect) and also shows the distribution within each condition (we see that items have a lot of variance on the horizontal axis but not on the vertical axis). However, this type of plot is not commonly used and thus may be more difficult for many readers to interpret. Furthermore, it is difficult now to plot the mean and confidence interval of the critical difference between conditions, as this is now represented on a diagonal axis (not shown) and would require trigonometry to understand.

Each of these plots has advantages and disadvantages; which one is most appropriate will depend on the context, the inferences the researcher is interested in making, and the message they are interested in communicating about the data. Note that in two of these three plots we have also supplemented the data aggregates with other model parameters to assist with drawing inferences from the plots. All of these model parameters, of course, have their limitations: the individual data points are aggregated over items or participants, the model estimates and confidence intervals are susceptible to influence by details of model specification, etc. They also each have their benefits: for example, since participant-wise or item-wise confidence intervals could yield inaccurate or conflicting statistical inferences (see Figure 5), the intervals based on an LMEM provide more accurate statistical inferences, while the individual participant and item means overlain provide a more accurate estimate of how the data are distributed around the model estimates.

Of course, there are many possible variations of plots like these. For instance, in this example we have combined by-participant and by-item means in the same plot. This provides a fuller picture than separate plots, and facilitates comparison between by-participant and by-item effects, but makes for a visualization that is busier and harder to read. Another alternative would be to plot by-participant and by-item aggregates separately, as in Figure 4, Figure 5, and the middle of Figure 10; such plots are cleaner and easier to read, but require more space.

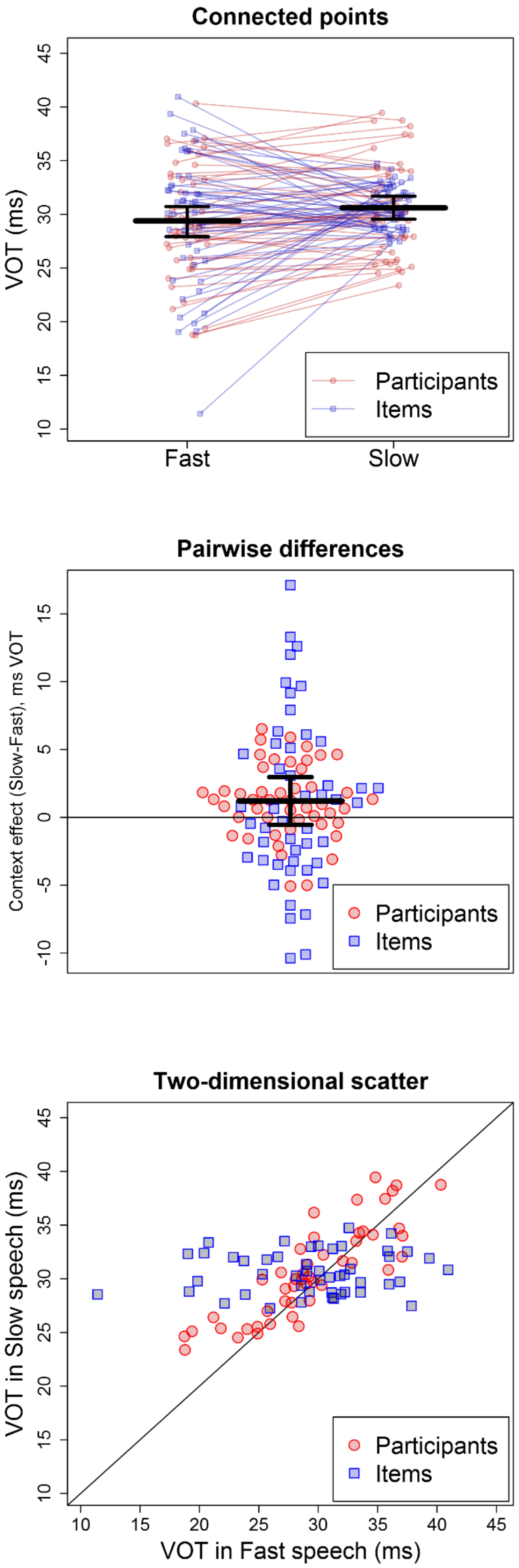
Connected points
VOT (ms)
10
15
20
25
30
35
40
45
Fast
Slow
Participants
Items
Pairwise differences
Context effect (Slow-Fast), ms VOT
-10
-5
0
5
10
15
Participants
Items
Two-dimensional scatter
VOT in Slow speech (ms)
10
15
20
25
30
35
40
45
VOT in Fast speech (ms)
10
15
20
25
30
35
40
45
Participants
Items

Figure 11. Three different visualizations of the same data. See text for details. In the top plot, thick horizontal lines represent the means for each condition, and error bars represent intervals for determining non-significant comparisons (Politzer-Ahles, 2017; see Footnote 2); the interpretation of the intervals is that when one condition's interval does not include another condition's mean and vice versa, those conditions are likely to be significantly different. In the middle plot, the thick horizontal line represents the fixed-effect estimate for the effect of speech rate in an LMEM, and the error bar represents the confidence interval of that difference (the standard error from the model output, times the critical z value for two-tailed α=0.05). The thin horizontal line in the middle plot and the thin diagonal line in the bottom plot both represent zero difference between conditions.

4.3 Simpler plots for communication of results

As discussed in section 3 with respect to Figure 10, sometimes a simpler plot is necessary for the sake of communicating the general pattern of data quickly. Likewise, it may be necessary to show several conditions together, rather than a series of pairwise differences, in order to illustrate a larger trend. In such cases, visualizations like the top or bottom plot in Figure 10 may be needed.

4.4 Detailed plots and/or raw data for posterity

Regardless of what simplified sort of visualization is eventually presented to clearly illustrate the comparisons of greatest interest for the intended research question, some more detailed representation of the data should be made available for others in the future to explore aspects of the dataset that were not shown in a simple representation. This could take the form of plots like those in Figure 11 and the bottom section of Figure 10 that are statistically informative, but might not be appropriate in a brief presentation, and/or highly detailed visualizations like Figure 6. Even if such plots are not necessary for addressing the primary comparison of interest, they will help others evaluate the conclusions of the research and may even lead to new hypotheses, as in the case of the eye-movement data mentioned above that led Staub and colleagues (2010, 2013) to test whether different aspects of the data distribution are affected by different manipulations.

Ultimately, the best way to address the issues described in this paper is to share the full dataset openly. As described above, no individual plot, no matter how detailed, can represent all the important dependencies in datasets with multiple random effects; therefore, the only way anybody can inspect these is to have the data available. Furthermore, for complicated designs with many potential comparisons, it is unfeasible to visualize all of them at once; by making the data public, a researcher gives others the chance to explore and plot comparisons that could not be included in published reports.

5 Conclusion

In this paper we have argued that the traditional practice of plotting aggregated data as bar plots with standard error bars is not always appropriate for most phonetic experiments. Furthermore, we have attempted to show that commonly-suggested solutions to the problem are not directly applicable when the data come from repeated measures designs with crossed random effects, because any method that shows the repeated nature of observations also requires some kind of aggregation; in such a case, aggregate data obscure aspects of the distribution and generally do not reflect the statistical model used for analysis. Data visualization must take into account the fact that multiple data points come from both a single individual and a single item. We have attempted to provide an outline for finding appropriate visualizations for a given dataset, taking into account these concerns and the outputs of the statistical models themselves. Most importantly, we encourage researchers to plot data in different ways at multiple steps throughout analysis and dissemination.

While we think these suggestions are a good start, we have mainly focused on the simplest possible cases, and other issues may also need to be considered when graphing other types of data. For instance, binomial data pose a special challenge to researchers who want to make plots integrating model parameters and rich data visualizations. The data themselves do not pose much special challenge relative to other data we have considered – visualizing binomial data generally requires some sort of aggregation (e.g., the proportion of voiced tokens or correct identifications for each participant or each item), which is no different than the need for aggregation in the VOT dataset we have discussed above. However, if one wants to integrate confidence intervals, for example, into the plot, these should come from a generalized mixed-effects model (Jaeger, 2008), in which the model parameters are expressed in terms of the logit of the proportion rather than the proportion itself (Tabachnik & Fidell, 2007). Thus, these parameters or confidence intervals must be expressed as proportions (using, e.g., the inverse logit function) before being added to a plot. (Alternatively, binomial data could be plotted in logit space rather than proportion space, but this generally would make them more difficult to understand—in phonetics and psycholinguistics it seems more common practice to express binomial data as proportions, i.e., as "75% accurate", rather than as logit or odds, i.e., as "correct responses were 3 times as likely as incorrect responses".) Another case we have not considered is continuous or ordinal independent variables. For instance, if people perform not just fast and slow speech, but four different speech rates, plots of connected observations (such as the top portion of Figure 11) become much more visually complex. On the other hand, "pairwise difference" sorts of plots are straightforwardly extended to such designs: a participant-wise or item-wise slope is a summary statistic just like a participant-wise or item-wise difference is, and can be plotted as a single point just like the pairwise differences in the middle portion of Figure 11.

In this paper we have discussed and provided examples for several issues in and approaches to visualization, including the following:

- How aggregates might obscure data distributions (Fig. 1) and pairs of repeated observations (Fig. 3)
- How error bars can be misleading (Figs. 2-3)
- How aggregation by participants or items can lead to different conclusions (Figs. 4-5)
- How plots of mixed-effect model coefficients are sensitive to model specification (Fig. 7) and do not directly reflect the observed data for each participant or item (Fig. 8)
- How plotting pairwise comparisons is complicated when a design includes polytomous variables (Figs. 9-10)
- Plots of by-participant and/or by-item aggregates (Figs. 4, 5, 7, 8, 9, 10, 11) and small multiple plots of each participant's and item's full dataset without aggregation (Fig. 6)

It is important to remember that there is not a one-size-fits-all approach. Different experiments will require different types of initial and final plots depending on the planned statistical analyses. We hope that the steps outlined here can serve as a guide for researchers as they try to determine what are the best visualizations for their data. We also emphasize that open data and code are necessary to allow researchers to understand and evaluate datasets and empirical claims.

Acknowledgements

This work was supported by ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*. We thank Kelly Berkson for allowing us to use her dataset in Figure 7 to illustrate an issue with BLUPs.

Reference List

Anscombe, F. (1973). Graphs in statistical analysis. *American Statistician, 27*, 17-21. DOI: 10.1080/00031305.1973.10478966

Baayen, R. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Cambridge University Press

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412. DOI: 10.1016/j.jml.2007.12.005

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavioural Research, 44*, 158-175. DOI: 10.3758/s13428-011-0123-7

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language, 68*, 255-278. DOI: 10.1016/j.jml.2012.11.001

Berkson, K., & Flego, S. (2017). Sonority sequencing effects in English: the emergence of the unmarked in a generally permissive language. https://osf.io/dv9hb/; DOI: 10.17605/OSF.IO/DV9HB

Blouin, D., & Rioppele, A. (2005). On confidence intervals for within-subjects designs. *Psychological Methods, 10*, 397-412. DOI: 10.1037/1082-989X.10.4.397

Chang, Y., & Lane, D. (2016). Generalizing across stimuli as well as subjects: a non-mathematical tutorial on mixed-effects models. *The Quantitative Methods for Psychology, 12*, 201-219.

Clark, H. (1973). The language as a fixed-effect fallacy: a critique of language and statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359. DOI: 10.1016/S0022-5371(73)80014-3

Eklund, A. (2016). beeswarm: the bee swarm plot, an alternative to stripchart. R package version 0.2.3. https://CRAN.R-project.org/package=beeswarm

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606.

Jaeger, T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446.

Judd, C., Westfall, J., & Kenny, D. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54-69. DOI: 10.1037/a0028347

Kliegl, R. (2014). Reduction of complexity of linear mixed models with double-bar syntax. *Rpubs*. https://rpubs.com/Reinhold/22193

Loftus, G., & Masson, M. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review, 1*, 476-490. DOI: 10.3758/BF03210951

Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *International Statistical Review, 25*, 7–22.

Politzer-Ahles, S. (2017). An extension of within-subject confidence intervals to models with crossed random effects. *The Quantitative Methods for Psychology, 13*, 75-94.

Politzer-Ahles, S., & Husband, E. (2018). Eye movement evidence for context-sensitive derivation of scalar inferences. *Collabra: Psychology, 4*, 3.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rousselet, G. (2017). Trimmed means. *Garstats*, https://garstats.wordpress.com/2017/11/28/trimmed-means/

Rousselet, G., Foxe, J., & Bolam, J. (2016). A few simple steps to improve the description of group results in neuroscience. *European Journal of Neuroscience, 44*, 2647-2651. DOI: 10.1111/ejn.13400

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

Staub, A., & Benatar, A. (2013). Individual differences in fixation duration distributions in reading. *Psychonomic Bulletin & Review, 20*, 1304-1311.

Staub, A., White, S., Drieghe, D., Hollway, E., & Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance, 36*, 1280-1293.

Tabachnick, B., & Fidell, L. (2007). *Using Multivariate Statistics*. Pearson.

Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H., & Kievit, R. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632-638.

Weissgerber, T., Milic, N., Winham, S., & Garovic, V. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biology, 13*, e1002128. DOI: 10.1371/journal.pbio.1002128