Journal Title
XX(X):1–25

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Analytic choices for analyzing multidimensional behavior: Many analysts test hypotheses about human speech

Stefano Coretta*[1], Timo Roettger[2], Joseph V. Casillas[3]

Abstract
TBA

Keywords
crowdsourcing science, data analysis, scientific transparency, speech, acoustic analysis

Introduction

In order to effectively accumulate knowledge, science needs to (i) produce data that can be replicated using the original methods and (ii) arrive at robust conclusions substantiated by the data. In recent coordinated efforts to replicate published findings, scientific disciplines have uncovered

[1]Institute of Phonetics and Speech Processing, Ludwig-Maximilian University Munich, Germany.
[2]Department of Applied Things, University B, City B, Country B
[3]Very Important Stuff Committee, Institute C, City C, Country C

Corresponding author:
Stefano coretta.
Email: s.coretta@lmu.de

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

surprisingly low success rates (e.g., Collaboration 2015; Camerer et al. 2018) leading to what is now referred to as the *replication crisis*. Beyond the difficulties of replicating scientific findings, a growing body of evidence suggests that the theoretical conclusions drawn from data are often variable even when researchers have access to reliable data. The latter situation has been referred to as the *inference crisis* (Rotello et al. 2015; Starns et al. 2019) and is, among other things, rooted in the inherent flexibility of data analysis (often referred to as researcher degrees of freedom: Simmons et al. 2011; Gelman and Loken 2014). Data analysis involves many different steps, such as inspecting, organizing, transforming, and modeling the data, to name a few. Along the way, different methodological and analytic choices need to be made, all of which may influence the final interpretation of the data. These researcher degrees of freedom are both a blessing and a curse.

They are a blessing because they afford us the opportunity to look at nature from different angles, which, in turn, allows us to make important discoveries and generate new hypothesis (e.g., Box 1976; Tukey 1977; De Groot 2014). They are a curse because idiosyncratic choices can lead to categorically different interpretations, which eventually find their way into the publication record where they are taken for granted (Simmons et al. 2011). Recent projects have shown that the variability between different data analysts is vast and can lead independent researchers to draw different conclusions from the same data set (e.g., Silberzahn et al. 2018; Starns et al. 2019; Botvinik-Nezer et al. 2020a). These studies, however, might still underestimate the extent to which analysts vary because data analysis is not merely restricted to statistical inference. Human behavior is complex and offers many ways to be translated into numbers. This is particularly true for fields that draw conclusions about human behavior and cognition from multidimensional data like audio or video data. In fields working on human speech production, for example, researchers need to make numerous decisions about what to measure and how to measure it. This is not trivial, given the temporal extension of the acoustic signal and its complex structural composition. Not only can decisions about measuring the signal influence downstream decisions about statistical modeling, but statistical results or modeling issues can also lead researchers to go back and revise earlier decisions about the measuring process itself.

In this article, we will investigate the variability in analytic choices when many analyst teams analyze the same speech production data set, a process that involves both decisions regarding the operationalization of a complex signal and decisions regarding the statistical analysis. Specifically, we will report the impact of analytic decision making on research results obtained by a minimum of 12 teams who will gain access to the same set of acoustic recordings in order to answer the same research question.

Researcher degrees of freedom

Data analysis comes with many decisions, like how to measure a given phenomenon or behavior, which data to submit to statistical modeling and which to exclude in the final analysis, or what inferential decision making procedure to apply. This can be problematic because humans show cognitive biases that can lead to erroneous inferences. Humans filter the world in irrational ways (e.g., Tversky and Kahneman 1974), seeing coherent patterns in randomness (Brugger 2001), convincing themselves of the validity of prior expectations ("I knew it", Nickerson 1998), and perceiving events as being plausible in hindsight ("I knew it all along", Fischhoff 1975). In conjunction with an academic incentive system that rewards certain discovery processes more than others (Sterling 1959; Koole and Lakens 2012), we often find ourselves exploring many possible analytic pipelines, but only reporting a select few.

This issue is particularly amplified in fields in which the raw data lend themselves to many possible ways of being measured (Roettger 2019). Combined with a wide variety of methodological and theoretical traditions as well as varying levels of statistical training across subfields, the inherent flexibility of data analysis might lead to a vast plurality of analytic approaches that can lead to different scientific conclusions. Analytic flexibility has been widely discussed from a conceptual point of view (Simmons et al. 2011; Wagenmakers et al. 2012; Nosek and Lakens 2014) and in regard to its application in individual scientific fields (e.g. Wicherts et al. 2016; Charles et al. 2019; Roettger 2019). This notwithstanding, there are still many unknowns regarding the extent of analytic plurality in practice.

Consequently, it is likely that a good part of published papers present overconfident interpretations of data and statistical results based on idiosyncratic analytic strategies (e.g., Simmons et al. 2011; Gelman

and Loken 2014). These interpretations, and the conclusions that derive from them, are thus associated with an unknown degree of uncertainty (dependent on the strength of evidence provided) and with an unknown degree of generalisability (dependent on the chosen analysis). Moreover, the same data could lead to very different conclusions depending on the analytic path taken by the researcher. However, instead of being critically evaluated, scientific results often remain unchallenged in the publication record. Despite recent efforts to improve transparency and reproducibility (e.g. Miguel et al. 2014; Klein et al. 2018) and the advent of freely available and accessible infrastructures, such as those provided by the Open Science Framework (osf.io), critical re-analyses of published analytic strategies are still uncommon because data sharing remains rare (Wicherts et al. 2006).

Crowdsourcing alternative analyses

Recent collaborative attempts have started to shed light on how different analysts tackle the same data set and have revealed a large amount of variability. In a collaborative effort, Silberzahn et al. (2018) let twenty-nine independent analysis teams address the same research hypothesis. Analytic approaches and consequently the results varied widely between teams. Sixty-nine percent of the teams found support for the hypothesis, and 31% did not. Out of the 29 analytic strategies, there were 21 unique combinations of covariates. Importantly, the observed variability was neither predicted by the team's preconceptions about the phenomenon under investigation nor by peer ratings of the quality of their analyses. The authors results suggest that analytic plurality may be an inevitable byproduct of the scientific process and not necessarily driven by different levels of expertise nor bias.

Several other recent studies corroborated this analytic flexibility across different disciplines. Both Dutilh et al. (2019) ad Starns et al. (2019) investigated analysts' choices when inferring theoretical constructs based on the same data set using computational models.

Both studies revealed vastly different modeling strategies, even though scientific conclusions were similar across analysis teams (see also Parker et al. 2020, on analytic flexibility in ecology and Botvinik-Nezer et al. (2020b) on neuroimaging data). Bastiaansen et al. (2020) crowdsourced clinical recommendations based on analyses of an individual patient.

Results suggests that analysts differed substantially regarding decisions related to both the statistical analysis of the data and the theoretical rationale behind interpreting the statistical results.

While these studies attested a large degree of analytic flexibility with possibly impactful consequences, they only investigated analytic decisions related to inferential statistics or the architecture of computational models. In these studies the data sets were fixed and neither data collection nor measurement could be changed. However, in many fields the primary raw data are complex signals that need to be operationalized relative to a theoretically motivated research question. This is especially true in the social sciences, where the phenomenon under investigation corresponds to both observable and unobservable human behavior. Decisions about how to measure a theoretical construct related to such behavior or its underlying cognitive processes might interact with downstream decisions about statistical modeling and vice versa.

For instance, Flake and Fried (2020) discuss the cascading impact that different practices can have on psychometric research. The authors highlight, among others, the following degrees of freedom in the choice and development of measures: definition of the theoretical construct, justification of the selected measure, description of the measure and of how it maps onto the construct, response coding and related transformations, and post-hoc modifications to the chosen measure. Taken together, these aspects alone exponentially increase the combinations of possible analytic choices, and hence flexibility in research outcomes.

In social sciences concerned with communication, human behavior is sometimes measured as a complex visual or acoustic signal. This added complexity introduces further fork junctions in the research pipe-line, and in turn further increases analytic flexibility. The present study looks at experimentally elicited speech production data in order to understand how this flexibility manifests itself in a scenario where complex decision procedures are involved in the operationalization and measuring of complex signals.

Operationalizing speech

Research on speech is at the heart of the cognitive sciences, informing psychological models of language, categorization, and memory, guiding methods for diagnosis and therapy of speech disorders, and facilitating

advancement in automatic speech recognition and speech synthesis. One major challenge in the speech sciences is the mapping between communicative intentions (the unobserved behavior) and their physical manifestation (the observed behavior).

Speech is a complex signal that is characterized by structurally different acoustic landmarks distributed throughout different temporal domains. Thus, choosing how to measure a phenomenon of interest is an important and non-trivial analytic decision. Take for example the following sentence in (1).

(1) "I can't bear another meeting on Zoom."

Depending on the speaker's intention, this sentence can be said in different ways. If, for instance, the speaker is exhausted by all their meetings, the speaker might acoustically highlight the word *another* or *meeting* to contrast it with other, more pleasant activities. If, on the other hand, the speaker is just tired of video conferences, as opposed to say face-to-face meetings, they might acoustically highlight the word *Zoom*.

If we decide to compare the speech signal associated with these two intentions, how can we quantify the difference between them? Given their physical manifestation (speech), what do we measure and how do we measure it? Because of the continuous and transient nature of speech, identifying speech parameters and temporal domains within which to measure those parameters becomes a non-trivial task. Utterances stretch over several thousand milliseconds and contain different levels of linguistically relevant units such as phrases, words, syllables, and individual sounds. The researcher is thus confronted with a considerable number of parameters and combinations thereof to choose from.

Speech categories are inherently multidimensional and dynamic: they consist of clusters of parameters that are modulated over time. The acoustic parameters of one category are usually asynchronous, i.e. they appear at different time points in the unfolding signal, and overlap with parameters of other categories (e.g. Jongman et al. 2000; Lisker 1986; Summerfield 1981; Winter 2014). A classical example is the distinction between voiced and voiceless stops in English (i.e. /b/ and /p/ in *bear* vs *pear*). This lexical contrast is manifested by many acoustic features which can differ depending on several factors, such as position

of the consonant in the word and surrounding sounds (Lisker 1977). Furthermore, correlates of the contrast can even be found away from the consonant, in temporally distant speech units. For example, the initial /l/ of the English words *led* and *let* is affected by the voicing of the final consonant (/t, d/) (Hawkins and Nguyen 2004).

The multiplicity of phonetic cues grows exponentially if we look at larger temporal domains as is the case for suprasegmental aspects of speech. For example, studies investigating acoustic correlates of word stress (e.g. the difference between *ínsight* and *incíte*) have been using a wide variety of measurements, including temporal characteristics (duration of certain segments or sub-segmental intervals), spectral characteristics (intensity, formants, and spectral tilt), and measurements related to fundamental frequency (f0) (e.g., Gordon and Roettger 2017). Moving on to the expression of higher-level functions, like information structure and discourse pragmatics, relevant acoustic cues can be distributed throughout even larger domains, such as phrases and whole utterances (e.g., Ladd 2008). Differences in position, shape, and alignment of pitch modulations over multiple locations within a sentence are correlated with differences in discourse functions (e.g. Niebuhr et al., 2011). The latter can also be expressed by global vs local pitch modulations (Van Heuven et al. 2002), as well as acoustic information within the temporal or spectral domain (e.g., Van Heuven and Van Zanten 2005). Extra-linguistic information, like the speaker's intentions, levels of arousal or social identity, are also conveyed by broad-domain parameters, such as voice quality, rhythm, and pitch (Foulkes and Docherty 2006; Ogden 2004; White et al. 2009).

In short, when testing hypotheses on speech production data, researchers are faced with many choices and possibilities. The larger the functional domain (e.g. segments vs words vs utterances), the higher the number of conceivable operationalizations. For example, several decisions have to be made when comparing the two realization of the sentence in (1), one of which is intended to signal emphasis on *another* and one of which emphasizes *zoom*.

(2a) I can't bear *ANOTHER* meeting on zoom.

(2b) I can't bear another meeting on *ZOOM*.

Do we compare only the word *another* in (2a) and (2b), or also the word *zoom*? Do we measure utterance-wide acoustic profiles, whole words, or just the stressed syllables? Do we average across the chosen time domain or do we measure a specific point in time? Do we measure fundamental frequency, intensity, or something else?

When looking at phrase-level temporal domains, the number of possible analytic pipelines increases exponentially. This plurality of analytic paths is illustrated in Figure 1. Even if we have decided to compare fundamental frequency (f0) of only the word *another* across the two utterances, there are still many choices to be made, all of which need to be justified. For example, we could measure f0 at specific points in time like the onset of the window, the offset, or the midpoint. We could also measure the value or time of the minimum or maximum f0 value. We could summarize f0 across the entire window and extract the mean, median or standard deviation of f0, all of which have been used to analyze speech data (see Gordon and Roettger 2017). And the garden of forking paths does not stop here. Figure 1, illustrates one specific procedure to automatically calculate f0. However, another options would be to filter the audio signal, to smooth the extracted f0 signal, to remove values that substantially deviate from surrounding values or expectations, either manually or automatically, and so on.

These decisions are intended to be made prior to any statistical analysis, but are at times revised a posteriori (i.e. after data collection and/or preliminary analyses) in light of unforeseen or surprising outcomes. These myriads of possible decisions are exponentiated by those researcher degrees of freedom related to statistical analysis (e.g. Wicherts et al. 2016). In sum, speech data are complex physical signals that constitutes an as of yet unappreciated amount of analytic flexibility when choosing measures and operationalizing underlying theoretical constructs. The present paper probes this garden of forking paths in the analysis of speech. To assess the variability in data analysis pipelines across independent researchers, we will provide analytic teams with an experimentally elicited speech production data set. The study aimed at investigating whether speakers acoustically modify utterances to signal unexpected referring expressions. In the following, we will shortly introduce the research question of the original study.

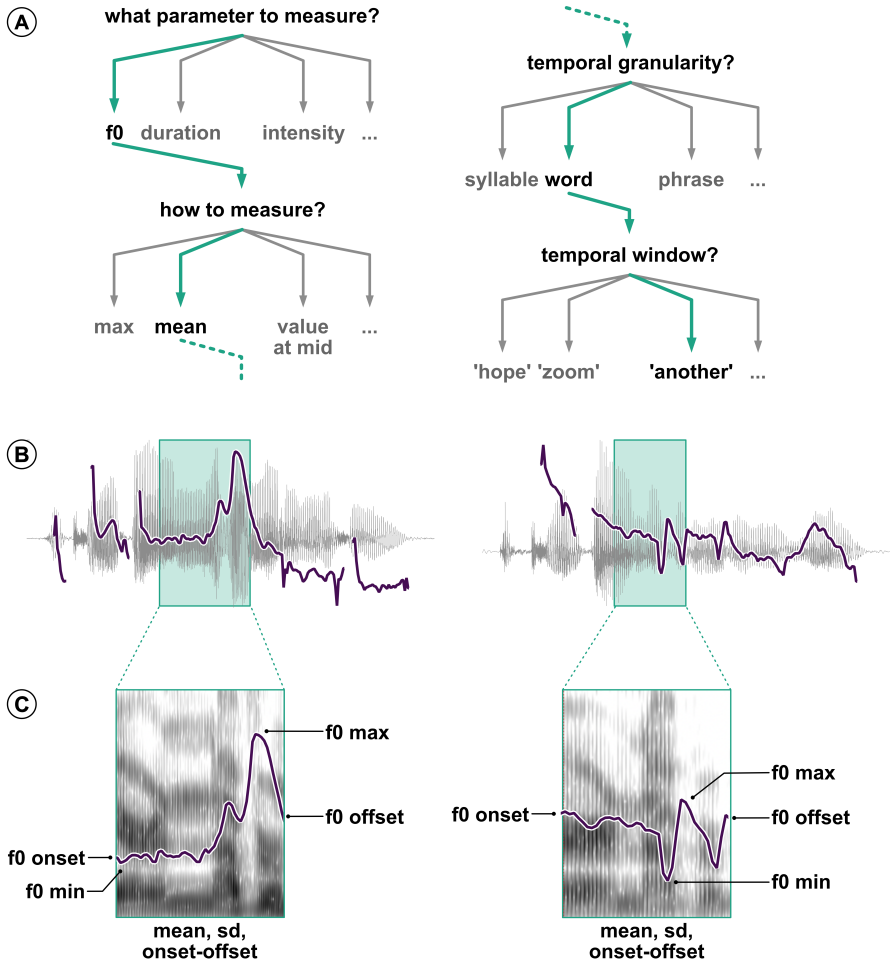

Figure 1. Illustrations of the analytic flexibility associated with acoustic analyses. (A) an example of multiple possible and justifiable decisions when comparing two utterances; (B) waveform and fundamental frequency (f0) track of two instances of utterances 2a and 2b. The words "another is highlighted by the green box"; (C) spectrogram and f0 track of the word another, exemplifying different operationalizations of differences in pitch.

The data set: The acoustic properties of atypical modifiers

Referring is one of the most basic and prevalent uses of language and one of the most widely researched areas in language science. It is an open question how speakers choose a referring expression when they want to

refer to a specific entity like a banana. The context within which an entity occurs (i.e., with other non-fruits, other fruits, or other bananas) plays a large part in determining the choice of referring expressions. Generally, speakers aim to be as informative as possible to uniquely establish reference to the intended object, but they are also resource-efficient in that they avoid redundancy (Grice 1975). Thus one would expect the use of a modifier, for example, only if it is necessary for disambiguation. For instance, one might use the adjective *yellow* to describe a banana in a situation in which there is a yellow and a less ripe green banana available, but not when there is only one banana to begin with.

Despite the coherent idea that speakers are both rational and efficient, there is much evidence that speakers are often over-informative: Speakers use referring expressions that are more specific than strictly necessary for the unambiguous identification of the intended referent (Sedivy 2003; Westerbeek et al. 2015; Rubio-Fernández 2016), which has been argued to facilitate object identification and making communication between speakers and listeners more efficient (Arts et al. 2011; Paraboni et al. 2007; Rubio-Fernández 2016). Recent findings suggest that the utility of a referring expression depends on how good it is for a listener (compared to other referring expressions) to identify a target object. For example, Degen et al. (2020) showed that modifiers that are less typical for a given referent (e.g. a blue banana) are more likely to be used in an over-informative scenario (e.g. when there is just one banana). This account, however, has mainly focused on content selection (Gatt et al. 2013), i.e. whether a certain referential expression is chosen or not, ignoring the fact that speech communication is much richer.

Even looking at morphosyntactically identical expressions, speakers can modulate utterances via suprasegmental acoustic properties like temporal and spectral modifications of the segments involved (e.g., Ladd 2008). Most prominently, languages use intonation to signal discourse relationships between referents (among other functions). Intonation marks discourse-relevant referents for being new or given information to guide the listeners' interpretation of incoming messages. Beyond structuring information relative to the discourse, a few studies suggested that speakers might use intonation to signal atypical lexical combinations [e.g. dimitrova2008prosodic; dimitrova2009did]. Referential expressions such as *blue banana* were produced with greater prosodic prominence

than more typical referent such as *yellow banana*. If true, these findings are in line with the idea of resource-efficient, rational language users who modulate their speech in order to facilitate listeners' comprehension process. However, the original conclusions of above studies were based on small (N = 10), unjustified sample sizes. Moreover, the data is not available, and the statistical analysis used anti-conservative co-variance structures, leaving reason to doubt the original conclusions.

To further illuminate the question of whether speakers modify speech to signal atypical referents, thirty native German speakers were recorded while interacting with a confederate (one of the experimenters) in a referential game. The participants had to verbally instruct the confederate to select a specified target object out of four object presented on a screen. Figure 2 shows the experimental procedure time-line. The subject and confederate were seated at the opposite sides of a table, each facing one of two mirrored computer screens. The participant and the experimenter could not see each other nor each others' screens. After a familiarization phase (see XXX), the subject first saw four colored objects in the top left, top right, bottom left, and bottom right corners of the screen. One of the objects served as the target, another as the competitor, and the remaining two objects served as distractors. Objects were referred to using noun phrases consisting of an adjective modifier denoting color and a modified object (e.g. *Gelbe Zitrone* 'yellow lemon', *Rote Gurke* 'red cucumber', *Rote Socken* 'red socks').

In the center of the screen, a black cube was displayed, which could be moved by the experimenter. The participant would read a sentence prompt out loud (*Du sollst den Würfel auf der ablegen* 'You have to put the cube on top of the ') to instruct the experimenter to drag the cube on top of one of the four depicted objects (the *competitor*) using the mouse. After the experimenter had moved the cube as instructed, the subject would read another sentence prompt (*Und jetzt sollst du den Würfel auf der ablegen* 'And now, you have to put the cube on top of the ') instructing the experimenter to move the cube on top of a different object (the *target*). The second utterance in the trial was the critical trial for analysis.

The two sentence prompts were used to create a focus contrast between the competitor and the target object. If the competitor and target objects differed but their color did not (e.g. *yellow banana* vs *yellow tomato*), the noun was discourse-pragmatically focused (the Noun Focus condition,

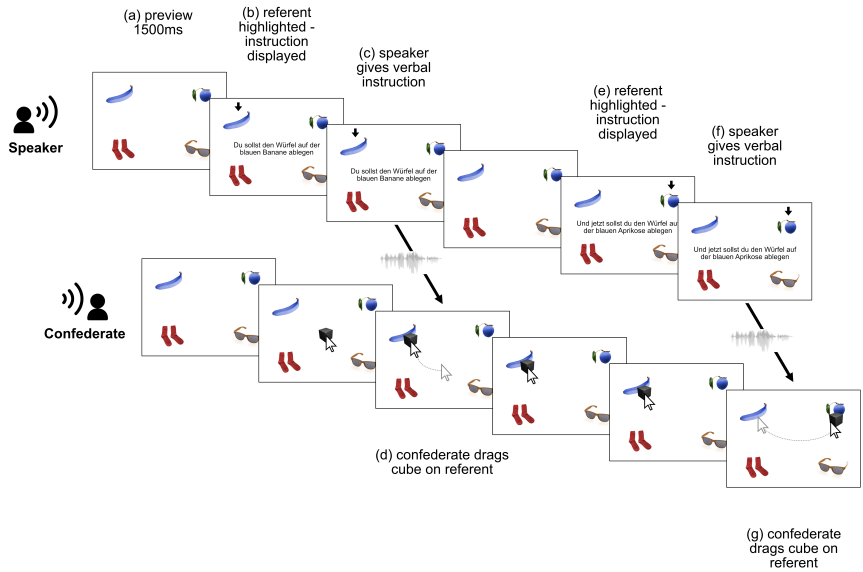

Figure 2. Experimental procedure. The upper row illustrates the trial sequence for the speaker (participant) and the lower row illustrates the trial sequence for the confederate. After a preview of 1500ms the speaker sees an arrow indicating one of the referent (b). Reading the orthographic instructions out loud, the speaker gives the confederate verbal instructions onto which referent they should drag the cube (c). The confederate, in turn, drags the black cube onto the target referent (d). Both the arrow and the orthographic instruction disappear from the speaker's screen and a new referent is indicated by an arrow on the same display alongside a new orthographic instruction (e). The speaker gives the confederate verbal instructions (f) which the confederate follows by dragging the cube onto the next referent (g).

NF). If the objects were the same but differed in color (e.g. *yellow banana* vs *blue banana*), the color adjective was discourse-pragmatically focused (the Adjective Focus condition, AF). If both the color and the object differed (e.g. *yellow banana* vs *blue tomato*), the whole noun phrase was discourse-pragmatically focused (the Adjective/Noun Focus condition, ANF). The NF condition constituted the experimentally relevant condition, while the AF and ANF conditions acted as fillers. The color-object combinations in the Noun Focus (NF) condition were manipulated with respect to their typicality. The combinations were either typical (e.g. *orange mandarin*), medium typical (e.g. *greem tomato*), or

atypical (e.g. *yellow cherry*), as established by a norming study that was conducted prior to the production experiment (see HERE*).

Each participant responded to 15 critical trials (NF condition). Each trial was repeated twice, yielding a total of 30 trials per participant and a grand-total 900 ($15 \times 2 \times 30$ participants) spoken utterances. All analysis teams will be given access to the entire data set and will be instructed to answer the following research question: *Do speakers acoustically modify utterances to signal atypical word combinations in referring expressions?*

Methods

As outlined in Section Operationalizing speech, researchers are faced with a large amount of analytic choices when analyzing a multidimensional signal such as speech. Analysts must identify and operationalize relevant measurements, as well as the temporal domain(s) from which these measurements are to be taken, and then possibly transform said measurements before submitting them to statistical models, which must be chosen alongside inferential criteria. The complexity of speech data constitutes the ideal testing ground to assess the upper bound of analytic flexibility that social science might face across disciplines. We will employ a meta-analytic approach to assess (i) the variability of the reported effects, and (ii) and how analytic and research-related predictors affect the researchers' final results.

We are closely following the data collection procedure proposed by Parker et al. (2020). The project involves the following five phases:

1. RECRUITMENT: We will recruit independent groups of researchers to analyze the data.
2. TEAM ANALYSIS: We will give researchers access to the speech corpus and let them analyze the data as they see fit.
3. REVIEW: We will ask reviewers to generate peer review ratings of the analyses based on methods (not results).
4. META-ANALYSIS: We will evaluate the variation among the different analyses.
5. WRITE-UP: Lastly, we will collaboratively produce the final manuscript.

*Temporary link: https://osf.io/5agn9/?view_only=3429acfd8eae4f0aafe2e17ad5b0983f

We estimate that this process, from the time of an in-principle acceptance of this Stage 1 Registered Report to the end of Phase 5, will take nine months. The factor most likely to delay our time line is the rate of completion of the original set of analyses by independent analysis teams.

Phase 1: Recruitment of analysts and initial survey

The initiating authors (SC, JC, TR) created an online landing page providing a general description of the project and a short pre-recorded slide show that summarizes the study and research question (https://many-speech-analyses.github.io). The project will be advertised via social media, using mailing lists for linguistic and psychological societies, and via word of mouth. The target population comprises active speech science researchers with a graduate/doctoral degree (or currently studying for a graduate/doctoral degree) in relevant disciplines. Researchers can choose to work independently or in a small team. For the sake of simplicity, we will refer to single researcher or small teams as ANALYSIS TEAMS. Recruitment for this project will commence upon receiving in-principle acceptance. As outlined above, our aim is to assess the variability of the reported effects and how these are affected by a set of predictors, rather than the meta-analytical estimate *per se*. Given the scarcity of previous meta-analytical studies that directly investigate researcher's degrees of freedom, as reviewed in the introduction, we are not able to establish a properly-grounded minimal sample size. Thus, we aim to recruit as many teams as possible, within the resource and time limits of the project.

We will simultaneously recruit volunteers to peer-review the analyses conducted by the analysis teams through the same channels. Our goal is to recruit a similar number of peer-reviewers and analysts, and to ask each peer reviewer to review a minimum of four analyses. If we are unable to recruit at least half the number of reviewers as analysis teams, we will ask analysts to serve also as reviewers (after they have completed their analyses).

All analysts and reviewers will share co-authorship on this manuscript and will participate in the collaborative process of producing the final manuscript. Informed consent will be obtained as part of the intake form.

Phase 2: Primary Data Analyses

The analysis teams will register and answer a demographic and expertise questionnaire (`intake_form.pdf`). PDF versions of all the questionnaires used in this study are available at LINK[†]. The questionnaire collects information on the analysts current position and self-estimated breadth and level of statistical expertise and acoustic analysis skills. We will then request that they answer the question: *Do speakers acoustically modify utterances to signal atypical word combinations?* To do so, they will be given the data generated by the experiment described in Section The data set. The data will include the audio recordings with corresponding time-aligned transcriptions in the form of Praat TextGrid files (LINK[‡]).

Once their analysis is complete, they will answer a structured questionnaire (`analytical_quest.pdf`), providing information about their analysis technique, an explanation of their analytic choices, their quantitative results, and a statement describing their conclusions. They will also upload their analysis files (including the additionally derived data and text files that were used to extract and pre-process the acoustic data), their analysis code (if applicable), and a detailed journal-ready statistical methods section.

Phase 3: Peer Review of Analyses

At a minimum, each analysis will be evaluated by four different reviewers, and each volunteer peer-reviewer will be randomly assigned to analyses from at least four analyst teams (the exact number will depend on the number of analysis teams and peer reviewers recruited). Each peer reviewer will register and answer a demographic and expertise questionnaire identical to that asked of the analysts. Reviewers will evaluate the methods of each of their assigned analyses one at a time in a sequence determined by the initiating authors. The sequences will be systematically assigned so that, if possible, each analysis is allocated to each position in the sequence for at least one reviewer.

The process for a single reviewer will be as follows. First, the reviewer will receive a description of the methods of a single analysis. This will

[†]Temporary link: https://osf.io/h6z8w/?view_only=3457f2b01a5f4435b3d917eb7b5134e8.
[‡]Temporary link: https://osf.io/5agn9/?view_only=3429acfd8eae4f0aafe2e17ad5b0983f

include the narrative methods section, the analysis team's answers to our survey questions regarding their methods, including analysis code and the data set. The reviewer will then be asked in an online questionnaire (`peer_review_quest.pdf`) to rate both the acoustic and the statistical analyses and to provide an overall rating, using a scale of 0-100, respectively. To help reviewers calibrate their rating, they will be given the following guidelines:

- 100. A perfect analysis with no conceivable improvements from the reviewer.
- 75. An imperfect analysis but the needed changes are unlikely to dramatically alter final interpretation.
- 50. A flawed analysis likely to produce either an unreliable estimate of the relationship or an over-precise estimate of uncertainty.
- 25. A flawed analysis likely to produce an unreliable estimate of the relationship and an over-precise estimate of uncertainty.
- 0. A dangerously misleading analysis, certain to produce both an estimate that is wrong and a substantially over-precise estimate of uncertainty that places undue confidence in the incorrect estimate.

The reviewers will also be given the option to include further comments in a text box for each of the three ratings.

After submitting the review, a methods section from a second analysis will then be made available to the reviewer. This same sequence will be followed until all analyses allocated to a given reviewer have been provided and reviewed. After providing the final review, the reviewer will be simultaneously presented with all four (or more) methods sections that the reviewer has just completed reviewing, the option to revise their original ratings, and a text box to provide an explanation.

Phase 4: Evaluating variation

Th initiating authors (SC, JC, TR) will conduct the analyses outlined in this section.

Descriptive statistics We will calculate summary statistics describing variation among analyses, including (a) the nature and number of acoustic measures (e.g. f0 or duration), (b) the operationalization and the temporal domain of measurement (e.g. mean of an interval or value at specified

point in time), (c) the nature and number of model parameters for both fixed and random effects (if applicable), (d) the nature and reasoning behind inferential assessments (e.g. dichotomous decision based on *p*-values, ordinal decision based on a Bayes factor), as well as the (e) mean, (f) standard deviation and (g) range of the standardizes effect sizes (see the next section for the standardization procedure).

Meta-analytic estimation We will investigate the variability in REPORTED EFFECT SIZES using Bayesian meta-analytic techniques. As the measure of variability, we will take the meta-analytic GROUP-LEVEL STANDARD DEVIATION (σ_{α_t}, see below), where group refers to the analytical teams.

Based on the common practices currently in place within the field, we anticipate that researchers will use multi-level/hierarchical/random-effects regression models, thus common measurements of effect size, such as Cohen's d, may be inappropriate. Since the variables used by the analysis teams might substantially differ in their measurement scales (e.g, Hertz for frequency vs milliseconds for duration), we will standardize all reported effects by refitting each REPORTED MODEL with centered and scaled continuous variables (z-scores, i.e. the observed values subtracted from the mean divided by the standard deviation) and sum-coded factor variables. Factor-level ordering for each categorical variable will mirror that of the original analyses. Each STANDARDIZED MODEL will be fitted as a Bayesian regression model with Stan (Team 2021), RStan (Team 2020), and brms (Bürkner 2017) in R (R Core Team 2020). For those reported models that were originally fitted within a frequentist approach, uniform distributions will be used as the priors of all parameters (with the restriction that only positive numbers will be included for scale parameters), making the standardized models in fact equivalent to the reported frequentist models. If a team has fitted Bayesian models, the same priors as reported by the team will be used in fitting the respective standardized model.

The coefficients of the critical predictors (i.e. critical according to the analysis teams' self-reported inferential criteria) obtained from the standardized models will be used as the STANDARDIZED EFFECT SIZE (η_i) of each reported model. If multiple predictors within a single analysis have been reported as critical, each will be included in the meta-analytic model (described in details in the next paragraph). Moreover, to account

for the differing degree of uncertainty around each standardized effect size, we will use the standard deviation of each effect size returned by the standardized models as the STANDARDIZED STANDARD ERROR (se_i). This will enable us to fit a so-called "measurement-error" model, in which both the standardized effect sizes and their respective standard errors are entered in the meta-analytic model. As a desired consequence, effect sizes with a greater standard error will be weighted less than those with a smaller standard error in the meta-analytic calculations.

After having obtained the standardized effect sizes η_i with related standard errors se_i, for each critical predictor of the individual reported analyses, we will fit a BAYESIAN RANDOM-EFFECTS META-ANALYSIS using a multilevel (intercept-only) regression model. The outcome variable will be the set of standardized effect sizes η_i. The likelihood of η_i is assumed to correspond to a normal distribution (Knight 2000). The analysis teams will be entered as a group-level effect (i.e., random effect, `(1 | team)`). The standard errors se_i will be included as the standard deviation σ_i of η_i to fit a measurement-error model, as discussed above. We will use regularizing weakly-informative priors for the intercept α ($Normal(0, 1)$) and for the group-level standard deviation σ_{α_t} ($HalfCauchy(0, 1)$). We will fit this model with 4 chains of Hamiltonian Monte-Carlo sampling for the estimation of the joint posterior distribution, using the No U-Turn Sampler (NUTS) as implemented in Stan (Team 2021), and 4000 iterations (2000 for warm-up) per chain, distributed across 4 processing cores. In case of divergent transitions, we will increase `adapt_delta`, `tree_depth`, and the number of iterations in this order until we obtain a model fit with no divergent transitions. The analysis will be conducted in R (R Core Team 2020) and fit using Stan (Team 2021), RStan (Team 2020), and brms (Bürkner 2017). The code used to run the model can be found here: INSERT LINK.

The posterior distribution of the population-level intercept α will allow us to estimate the range of probable values of the standardized effect size $\hat{\eta}$. The posterior distribution will further allow us to investigate the effect of a set of analytic and researcher-related predictors, detailed in the next section. Crucially, the posterior distribution of σ_{α_t} (the standard deviation of the group-level effect of team) will allow us to quantify the degree of variation between teams on a standardized scale.

Analytic and researcher-related predictors affecting effect sizes As a second step, we will investigate the extent to which the individual standardized effect sizes are affected by a series of ANALYTIC AND RESEARCHER-RELATED PREDICTORS.

Analytic predictors. We will estimate the influence of the following predictors related to the analytic characteristics of each team's reported analysis:

- *Measure of uniqueness* of individual analyses for the set of predictors in each model [numeric].
- *Measure of conservativeness* of the model specification, as the number of random/group-level effects included [numeric].
- *Number of post-hoc changes to the acoustic measurements* the teams will report to have carried out [numeric].
- *Number of models* the teams will report to have run [numeric].
- *Major dimension* that has been measured to answer the research question [categorical].
- *Temporal window* that the measurement is taken over [categorical].
- *Average peer-review rating*, as the mean of the overall peer-review ratings for each analysis [numeric].

Following Parker et al. (2020), the measure of uniqueness of predictors will be assessed by the Sørensen-Dice Index (SDI, Dice 1945; Sørensen 1948). The SDI is an index typically used in ecology research to compare species composition across sites. For our purposes, we will treat predictors as species and individual analyses as sites. For each pair of analyses (X, Y), the SDI will be obtained using the following formula:

$$\text{SDI} = \frac{2|X \cap Y|}{|X| + |Y|}$$

where $|X \cap Y|$ is the number of variables common to both models in the pair, and $|X| + |Y|$ is the sum of the number of variables that occur in each model.

In order to generate a unique SDI for each analysis team, we will calculate the average of all pairwise SDIs for all pairs of analyses using the `beta.pair()` function in the betapart R package (Baselga et al. 2020).

The major measurement dimension of each analysis will be categorized according to the following possible groups: *duration*, *amplitude*,

fundamental frequency, *other spectral properties* (e.g. frequency, center of gravity, harmonics difference, etc.), *other measures* (e.g. principal components, vowel dispersion, etc.). The temporal window that the measurement is taken over is defined by the target linguistic unit. We assume the following relevant linguistic units: *segment*, *syllable*, *word*, *phrase*. Since each analysis will receive more than one peer-review rating, we will calculate the mean rating and its standard deviation for each. These will be entered in the model formula as a measurement-error term (`me(mean, sd)` in brms).

Researcher-related factors. We will also include the following predictors:

- *Research experience* as the elapsed time from receiving the PhD. Negative values will indicate that the person is a student or graduate studens [numeric].
- *Initial belief* in the presence of an effect of atypical noun-adjective pairs on acoustics, as answered during the intake questionnaire [numeric].

Since we will obtain scores for each member in the analytic teams, we will calculate the team mean and standard deviation for each predictor. These will be entered in the model formula as a measurement-error term (`me(mean, sd)` in brms).

Model specification. The model will be fitted as a measurement-error model, with the predictors detailed in the preceding paragraphs. The outcome variable of the model will be the standardized effect sizes and related standard deviation.

A normal distribution will be used as the likelihood function of $\alpha_{t[i]}$. The mean of $\alpha_{t[i]}$ is modeled on the basis of the overall intercept β and on the coefficients of each predictor. The numeric predictors will be centered and scaled and the categorical predictors will be sum coded. As the prior for the intercept and the predictors we will use a normal distribution with mean 0 and standard deviation 1. The model will be run with the same settings as with the meta-analytic model. The code used to run the model can be found here: INSERT LINK.

Data management All relevant data, code, and materials will be publicly archived on the Open Science Framework (https://osf.io/3bmcp/). Archived data will include the original data sets distributed

to all analysts, any edited versions of the data analyzed by individual groups, and the data we analyze with our meta-analyses, which include the effect sizes derived from separate analyses, the statistics describing variation in model structure among analysis teams, and the anonymized answers to our questionnaires of analysts and peer reviewers. Similarly, we will archive both the analysis code used for each individual analysis and the code from our meta-analyses. We will also archive copies of our survey instruments from analysts and peer reviewers.

We will exclude from our synthesis any individual analysis submitted after we have completed peer review or those unaccompanied by analysis files that allow us to understand what the analysts did. We will also exclude any individual analysis that does not produce an outcome that can be interpreted as an answer to our primary question.

Phase 6: Collaborative Write-Up of Manuscript

Analysts and initiating authors will discuss the limitations, results, and implications of the study and collaborate on writing the final manuscript for review as a stage-2 Registered Report.

Glossary

- **Analysis team**: team of analysts or single analyst.
- **Reported effect sizes**: effect sizes reported by each analysis team.
- **Standardized model**: Bayesian refit of the team's model.
- **Standardized effect sizes**: (η_i) effect sizes returned by the standardised models.
- **Standardized standard error**: (se_i) standard deviation of the standardized effect sizes.
- **Bayesian random-effects meta-analysis** and **meta-analytic model**: multilevel intercept-only regression model for meta-analysis.
- **Meta-analytic group-level standard deviation**: (σ_{α_t}) group-level standard deviation returned by the meta-analytic model.
- **Analytic and researcher-related predictors**: predictors used in model XX.

References

Arts A, Maes A, Noordman LG and Jansen C (2011) Overspecification in written instruction. *Linguistics* 49(3): 555–574.

Baselga A, Orme D, Villeger S, De Bortoli J, Leprieur F and Logez M (2020) betapart: Partitioning beta diversity into turnover and nestedness components. URL https://CRAN.R-project.org/package=betapart. R package version 1.5.2.

Bastiaansen JA, Kunkels YK, Blaauw FJ, Boker SM, Ceulemans E, Chen M, Chow SM, de Jonge P, Emerencia AC, Epskamp S et al. (2020) Time to get personal? the impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research* 137: 110211.

Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, Kirchler M, Iwanir R, Mumford JA, Adcock RA et al. (2020a) Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* : 1–7.

Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, Kirchler M, Iwanir R, Mumford JA, Adcock RA et al. (2020b) Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582(7810): 84–88.

Box GE (1976) Science and statistics. *Journal of the American Statistical Association* 71(356): 791–799.

Brugger P (2001) From haunted brain to haunted science: A cognitive neuroscience view of paranormal and pseudoscientific thought. *Hauntings and Poltergeists: Multidisciplinary Perspectives* : 195–213.

Bürkner PC (2017) brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1): 1–28. DOI:10.18637/jss.v080.i01.

Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, Kirchler M, Nave G, Nosek BA, Pfeiffer T et al. (2018) Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour* 2(9): 637–644. DOI: 10.1038/s41562-018-0399-z.

Charles SJ, Bartlett JE, Messick KJ, Coleman TJ and Uzdavines A (2019) Researcher degrees of freedom in the psychology of religion. *The International Journal for the Psychology of Religion* 29(4): 230–245.

Collaboration OS (2015) Estimating the reproducibility of psychological science. *Science* 349(6251). DOI:10.1126/science.aac4716.

De Groot AD (2014) *Thought and choice in chess*, volume 4. Walter de Gruyter GmbH & Co KG.

Degen J, Hawkins RD, Graf C, Kreiss E and Goodman ND (2020) When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review* .

Dice LR (1945) Measures of the amount of ecologic association between species. *Ecology* 26(3): 297–302.

Dutilh G, Annis J, Brown SD, Cassey P, Evans NJ, Grasman R, Hawkins GE, Heathcote A, Holmes WR, Krypotos AM et al. (2019) The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review* 26(4): 1051–1069.

Fischhoff B (1975) Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human perception and performance* 1(3): 288.

Flake JK and Fried EI (2020) Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science* 3(4): 456–465. DOI:10.1177/2515245920952393.

Foulkes P and Docherty G (2006) The social life of phonetics and phonology. *Journal of Phonetics* 34(4): 409–438.

Gatt A, van Gompel RP, van Deemter K and Krahmer E (2013) Are we bayesian referring expression generators. Cognitive Science Society.

Gelman A and Loken E (2014) The statistical crisis in science: data-dependent analysis–a" garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American scientist* 102(6): 460–466.

Gordon M and Roettger T (2017) Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard* 3(1): 1–11.

Grice HP (1975) Logic and conversation. In: *Speech acts*. Brill, pp. 41–58.

Hawkins S and Nguyen N (2004) Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *Journal of Phonetics* 32(2): 199–231.

Jongman A, Wayland R and Wong S (2000) Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America* 108(3): 1252–1263.

Klein O, Hardwicke TE, Aust F, Breuer J, Danielsson H, Mohr AH, IJzerman H, Nilsonne G, Vanpaemel W, Frank MC et al. (2018) A practical guide for transparency in psychological science. *Collabra: Psychology* 4(1).

Knight K (2000) *Mathematical Statistics*. Chapman and Hall, New York.

Koole SL and Lakens D (2012) Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science* 7(6): 608–614.

Ladd DR (2008) *Intonational phonology*. Cambridge University Press.

Lisker L (1977) Rapid versus rabid: A catalogue of acoustic features that may cue the distinction. *The Journal of the Acoustical Society of America* 62(S1): S77–S78.

Lisker L (1986) "voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and speech* 29(1): 3–11.

Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, Gerber A, Glennerster R, Green DP, Humphreys M, Imbens G et al. (2014) Promoting transparency in social science research. *Science* 343(6166): 30–31.

Nickerson RS (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2(2): 175–220.

Nosek BA and Lakens D (2014) A method to increase the credibility of published results. *Social Psychology* 45(3): 137–141.

Ogden R (2004) Non-modal voice quality and turn-taking in Finnish. *Sound patterns in interaction: cross-linguistic studies from conversation* : 29–62.

Paraboni I, Van Deemter K and Masthoff J (2007) Generating referring expressions: Making referents easy to identify. *Computational linguistics* 33(2): 229–254.

Parker T, Fraser H, Nakagawa S, Gould EB, Griffith S, Vesk P and Fidler F (2020) Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology [passed peer review and granted in-principle

acceptance March 2020]. *BMC Biology* DOI:10.6084/m9.figshare.12034833.v1. URL https://springernature.figshare.com/articles/journal%5Fcontribution/Same%5Fdata%5Fdifferent%5Fanalysts%5Fvariation%5Fin%5Feffect%5Fsizes%5Fdue%5Fto%5Fanalytical%5Fdecisions%5Fin%5Fecology%5Fand%5Fevolutionary%5Fbiology%5FRegistered%5FReport%5FStage%5F1%5FProtocol%5F/12034833.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roettger TB (2019) Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1).

Rotello CM, Heit E and Dubé C (2015) When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review* 22(4): 944–954.

Rubio-Fernández P (2016) How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in psychology* 7: 153.

Sedivy JC (2003) Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research* 32(1): 3–23.

Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, Bahník Š, Bai F, Bannard C, Bonnier E et al. (2018) Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science* 1(3): 337–356.

Simmons JP, Nelson LD and Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22(11): 1359–1366.

Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab* 5(4): 1–34.

Starns JJ, Cataldo AM, Rotello CM, Annis J, Aschenbrenner A, Bröder A, Cox G, Criss A, Curl RA, Dobbins IG et al. (2019) Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Advances in Methods and Practices in Psychological Science* 2(4): 335–349.

Sterling TD (1959) Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American statistical association* 54(285): 30–34.

Summerfield Q (1981) Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance* 7(5): 1074.

Team SD (2020) RStan: the R interface to Stan. URL http://mc-stan.org/. R package version 2.21.2.

Team SD (2021) Stan modeling language users guide and reference manual, v2.26.0. URL http://mc-stan.org/.

Tukey JW (1977) *Exploratory data analysis*, volume 2. Reading, MA.

Tversky A and Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *science* 185(4157): 1124–1131.

Van Heuven VJ, Haan J, Gussenhoven C and Warner N (2002) Temporal distribution of interrogativity markers in Dutch: A perceptual study. In: *Laboratory Phonology 7*, volume 4. Walter de Gruyter, p. 61.

Van Heuven VJ and Van Zanten E (2005) Speech rate as a secondary prosodic characteristic of polarity questions in three languages. *Speech Communication* 47(1-2): 87–99.

Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL and Kievit RA (2012) An agenda for purely confirmatory research. *Perspectives on Psychological Science* 7(6): 632–638.

Westerbeek H, Koolen R and Maes A (2015) Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in psychology* 6: 935.

White L, Payne E and Mattys SL (2009) Rhythmic and prosodic contrast in Venetan and Sicilian Italian. In: Vigario M, Frota S and Freitas MJ (eds.) *Phonetics and phonology: Interactions and interrelations*. Amsterdam: John Benjamins, pp. 137–158.

Wicherts JM, Borsboom D, Kats J and Molenaar D (2006) The poor availability of psychological research data for reanalysis. *American psychologist* 61(7): 726.

Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM and van Assen MALM (2016) Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology* 7. DOI:10.3389/fpsyg.2016.01832.

Winter B (2014) Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays* 36(10): 960–967.