How reliable are standard reading time analyses?

Hierarchical bootstrap reveals substantial power over-optimism and scale-dependent Type I error inflation

Zachary J. Burchill[1] & T. Florian Jaeger[1,2]

[1] Brain and Cognitive Sciences, University of Rochester, NY 14627, USA

[2] Computer Science, University of Rochester, NY 14627, USA

Contact:
T. Florian Jaeger
fjaeger@ur.rochester.edu
University of Rochester,
Meliora Hall,
Rochester, NY 14627, USA

Abstract

We investigate the statistical power and Type I error rate of the two most common approaches to reading time (RT) analyses: assuming normality of residuals and homogeneity of variance in raw or log-transformed RTs. We first show that the assumptions of such analyses—such as *t*-tests, ANOVAs, and linear mixed-effects models—are neither consistently met by raw RTs, nor by log-transformed RTs (or any other common power transforms, incl. inverse-transformed RTs). Only a non-power transform (log-shift) provides a decent fit for all data sets and data preparation steps we consider. We then compare the statistical power and Type I error rate for linear mixed-effects models over raw or log-transformed RTs. Previous studies on this matter relied on parametrically generated data. We show why this is problematic, and introduce as an alternative a hierarchical bootstrap approach over naturally distributed reading times. This approach yields substantially different—and arguably more informative—results than the parametric simulation approaches we compare it to. Our results suggests that it is time to heed the advice others have provided for reading research: for any but the simplest designs, we find both the rate of spurious significances and the rate of undetected true effects can *strongly* depend on the scale (e.g., raw or log-RTs) in which effects are assumed to be linear. Researchers should thus clearly motivate the choice of analysis based on theoretical grounds, assess the robustness of findings under different analysis approaches, and discuss potential mismatches between analyses. The R scripts and libraries shared in the accompanying OSF repo allow researchers to assess the reliability of their analyses via hierarchical bootstrap over their own data.

Keywords: reading times; data analysis; power; Type I error; hierarchical bootstrap

Introduction

Reading time data are widely used across the cognitive and neurosciences, medical contexts, and research on education. They have informed theories of language understanding, serve a key role in testing theories of reading, and lend insight to pedagogical contexts. However, many of the most common methods used to analyze reading times (RTs)—such as *t*-tests, analysis of variance (ANOVA), or linear mixed models (LMM)—assume that residual RTs are normally distributed with variance that is independent of the mean (homogeneity of variance). Both the normality and the homogeneity assumptions are likely to be false. Reading is a complex task that spans perceptual, cognitive, and motor processes. To the extent that the component processes are not perfectly information encapsulated, their contributions are *not* expected to be purely additive in raw RTs, making raw RTs unlikely to follow a normal distribution (Stephen & Mirman, 2010). RTs are also known to exhibit a soft lower bound and a positive skew, both of which are unexpected under the assumption of normality (similar properties are found for related psychometric data, such as reaction times in two-alternative forced choice tasks, Ratcliff & Smith, 2004; multi-choice tasks, Usher & McClelland, 2001; picture or color naming times, Heathcote, Popiel, & Mewhort, 1991; Snodgrass & Yuditsky, 1996).

These and related considerations have motivated a number of alternative approaches to RT analyses, ranging from relatively simple to increasingly complex. On the more advanced end, researchers have proposed time series and related models that correct for the lack of independence between temporally adjacent observations, capturing that each RT can reflect processing of not just the current input, but also preceding input ("spillover" analyses, Ehrlich & Rayner, 1983; Mitchell, 1984; generalized additive mixed models with auto-correlations, Baayen et al., 2016; or continuous-time deconvolutional regression, Shain & Schuler, 2021). Others have explored approaches that analyze RTs as a sum of several independent processes (e.g., log-shift, Ex-Gaussian, or other mixture models; see Nicenboim & Vasishth, 2018; Rouder, 2005; Staub & Benatar, 2013), or developed process models of reading that can be fit against RTs or eye-

movements during natural reading (e.g., SWIFT, Engbert et al., 2005; ACT-R, Lewis et al., 2013; Lewis & Vasishth, 2005; EZ-READER Reichle et al., 2003). Any of these more advanced approaches offers unique opportunities to researchers to better understand their data, and to increase the reliability of their analyses. They do, however, also come with challenges, such as additional computational complexity and the need for additional statistical training.

While the future of reading analyses likely lies in these more advanced approaches (for an excellent review, see Shain & Schuler, 2021), it does not appear that this future is immanent: the majority of reading research continues to employ analysis approaches that are computationally less demanding, and require less expertise to interpret. The most common of these simpler approaches employ variants of the linear model—such as *t*-tests, ANOVAs, or LMMs—over inverse- or log-transformed, rather than raw RTs. Recent informal surveys of the field suggest linear models over raw or power transformed RTs account for as much as 98% of RT analyses (Nicklin & Plonsky, 2020; see also Liceralde & Gordon, 2022). The same reviews suggest that analyses over raw, untransformed RTs remain the most common approach—despite the obvious problems with the assumptions they make—followed by analyses over log-transformed RTs. Given the continued prevalence of these approaches, the present work aims to shed light on how this choice affects the reliability—i.e., the Type I error rate and statistical power—of RT analyses. After all, ease of interpretability—sometimes evoked as an argument in favor of simpler approaches (e.g., Osborne, 2002)—is only of value if the interpretations drawn from the data are valid (see also Lo & Andrews, 2015). We thus assess whether one of the two approaches—analyses over raw or log-transformed RTs—is to be consistently preferred over the other in terms of Type I error rates and/or power. Inflated Type I error rates of either approach would call into question theories that are built on those analyses and findings. And inflated power estimates would lead researchers to be over-confident in their results, a matter that has only gained in relevance with an increasing focus on replicability. To further contextualize our results, we compare both approaches to a simple alternative described in more detail below, the log-shift transform.

The findings we present below differ in important ways from previous work, and we present evidence that this is likely due to common (unvalidated) parametric assumptions made in previous work about the

distribution of RTs. We find that neither LMMs over raw nor LMMs over log-transformed RTs offer a one-size-fits-all analysis solution. Both can lead to substantial Type I error inflation and wasted power, depending on the experiment's design. Indeed, there are reasons to believe that even more advanced methods like the ones described above are unlikely to completely resolve these issues. Our results do, however, suggest a pattern of *when* which of the two analysis approaches is to be preferred.

Based on our results, we estimate that a large proportion of RT studies might be underpowered, and that analyses with interactions additionally suffer from inflated Type I errors. Reading researchers should either employ simulation studies like ours to show that the issues we identify do not apply to *their* data, or demonstrate that their results are robust to at least the most common transformations. If the latter is not the case, researchers should clearly motivate how their theoretical assumptions justify the interpretation of their results, and/or carefully discuss why different analysis approaches yield different results for their data (see also Baayen et al., 2016; Staub, 2021, p. 15). Beyond improving the reliability of RT analyses, this will also facilitate comparison of results across studies, which suffers when different studies (even by the same authors—our own work included) employ different types of analyses.

The present study

We present four statistical simulation studies (complemented by extensive auxiliary studies in the supplementary information, SI). We focus on RTs in self-paced reading paradigms (SPR), though future studies could employ the hierarchical bootstrap approach we present to validate analysis approaches for eye-tracking during reading. SPR continues to be frequently used in reading research. It is inexpensive, easy to implement, and yields a single RT measure.

Study 1 begins by characterizing the distribution of RTs in three different SPR data sets. While these questions have received attention for simpler psychometric tasks (Baayen & Milin, 2010; Brysbaert & Stevens, 2018; Kliegl et al., 2010; Lachaud & Renaud, 2011; Lo & Andrews, 2015; Rouder, 2005; Rouder et al., 2005; Schramm & Rouder, 2019; Wagenmakers & Brown, 2007), it is by no means clear that an ability as complex as reading yields distributions that resemble those of simpler two-alternative forced-

choice tasks (see also Wagenmakers et al., 2005). Together, the three RT data sets span three common types of SPR experiments that psycholinguistic research draws on: factorial experiments conducted in the lab, factorial experiments conducted over the web via crowdsourcing, and studies conducted over reading corpora.

We first ascertain that RTs in all three data sets indeed violate the assumptions of normality and homogeneity. We then introduce the non-parametric hierarchical bootstrap approach employed in the remainder of the article. We use this approach to compare the distribution of bootstrapped natural RTs against the distribution of RTs that are parametrically generated under common power transformations. We find that none of the common transformations yields distributions that match those of natural RTs, though the log-transform provides a better fit than other common power transformations. This motivates the question we address in the remainder of the article: do the distributional properties of RTs have detrimental consequences for the reliability of the most common approaches to RT analyses? Not all unmet analysis assumptions have practical consequences on statistical power and Type I error rates. For example, while linear mixed models can be relatively robust to violations of normality (Knief & Forstmeier, 2018), heterogeneous variances are known to inflate Type I error rates for analyses of categorical data (e.g., if ANOVA or LMMs are used to analyze binomially distributed responses, Dixon, 2008; Jaeger, 2008).

Studies 2-4, as well as five auxiliary studies in the SI, address this question for the two power transformations (Box & Cox, 1964) that are most commonly employed in RT analyses: the identity transform (raw RTs) and the log-transform. Study 2 begins to investigate the consequence of naturally distributed RTs on power and Type I error rates in a simple by-2 design (a two-way manipulation, e.g., comparing treatment against control). Previous work has addressed this question for simpler psychometric paradigms (Brysbaert & Stevens, 2018; Lachaud & Renaud, 2011; Liceralde & Gordon, 2022; Ratcliff, 1993; Schramm & Rouder, 2019). These studies exclusively employed parametrically generated data. Liceralde & Gordon (2022), for example, fit LMMs to reaction time data, and then generate new reaction times from this parametric model (while adding the assumption that trial-level residuals follow a Gamma distribution). Type I and power analyses are conducted over these parametrically generated reaction times.

This raises questions about the extent to which the results from these studies generalize to *actual* (rather than parametrically generated) reaction time data, as collected in experiments. Study 2 begins to address this question for reading times by means of non-parametric hierarchical bootstrap. This approach avoids specific parametric assumptions about the distribution of RTs.

Study 3 confirms that the results of Study 1 would indeed differ if we had used parametrically generated data, instead of bootstrapped naturally distributed RTs. This offers an explanation as to why our results support different conclusions than previous work, and highlights the need to carefully consider the approach to data generation when evaluating statistical power and Type I errors. Study 3 also has far reaching consequences for reading research: we find that reading research might have *routinely* and *substantially* over-estimated statistical power (the issues we identify hold in *addition* to other wide-spread issues with power estimates, such as the "significance filter", Vasishth et al., 2018).

Finally, Study 4 moves beyond the simple by-2 design and tests how statistical tests of *interactions* are affected by assumptions about the distribution of RTs. Interactions, or a lack therefore, are often used to argue for or against theories. But interactions are also known to be particularly vulnerable to inadequate assumptions about the data. For example, a frequent finding is that conditions with overall slower RTs are more strongly affected by a manipulation than conditions with overall faster RTs. While such differences are routinely interpreted as meaningful, they can be the consequence of the soft lower bound of RTs: in conditions in which reading is already fast, it is difficult to detect further increases in reading speed. As we discuss as part of Study 4, reasoning about interactions is further complicated due to their “scale-dependence” (e.g., Loftus, 1978)—an issue that is conceptually independent of, but might in practice interact with, assumptions about the distribution of RTs (e.g., Lo & Andrews, 2015; Staub, 2021; Sternberg, 1969b, to which we return in the general discussion). These issues are likely *not* specific to linear model analyses, but rather are expected to extend to most advanced analyses (including all of the approaches mentioned further up). They are thus likely to persist even if the field eventually embraces those more advanced approaches.

Open Science Statement

The source data, simulation summaries, and the R code for our simulation studies are shared via OSF (https://osf.io/uymfp/). The three source data sets from which we bootstrapped contain a total of almost 1 million per-word RTs from over 400 subjects and 600 sentence items. The R code includes general code for both the hierarchical bootstrap and parametric generation of RTs, allowing researchers to conduct analyses similar to ours for their own data. The code affords flexible parallelization over multiple (local or remote) cores via R's *future* package (Bengtsson, 2019). We hope that this will help the field to jointly build a stronger understanding of how the distributional properties of RTs affect statistical power and Type I errors, and the extent to which these effects depend on the experimental paradigm.

Study 1

We begin by introducing the three source data sets we use in our simulation studies. We characterize the distribution of RTs in those data sets, and show that it exhibits the non-normal properties that have previously been documented for reaction time data from simpler psychometric tasks. We then introduce the two types of data generation procedures employed in the remainder of our simulation studies: a hierarchical bootstrap over the natural RTs and parametrically generated RTs that correspond to the assumptions implied by the power transformations that are most commonly applied in RT analyses (the identity, logarithmic, and reciprocal transform). We show that the hierarchical bootstrap generates data sets with statistical properties that closely resemble those of natural RTs in the three source data sets. However, none of the most common power transformations consistently matches those distributions. That said, some transformations come closer than others, and a comparatively simple adjustment to one of the common power transformations provides a decent fit against all three source data sets and data preparation steps we consider.

Beyond its primary goals, Study 1 aims to illustrate how assumptions baked into the data generation process can strongly affect the outcome of statistical simulations. Such assumptions typically receive relatively little attention in, e.g., power calculations. Yet, as we will see in Studies 2 and 3, they can have

far-reaching consequences. Like some of the other findings we report, this insight might not be particularly surprising to statistical experts or appear obvious in hindsight. We report and emphasize these points because we believe they remain under-appreciated in actual practice (incl. in our own previous work).

Source data

We employ three source data sets that reflect three common SPR paradigms used in psycholinguistic research: lab- and web-based experiments as well as reading corpora. This allows us to test to what extent the issues we identify in our studies are at least somewhat likely to generalize to other self-paced RT data, rather than being a consequence of specific paradigms (web- vs. lab-based; factorial design vs. reading corpus; reading of isolated sentences vs. connected narratives). Table 1 shows that the three data sets also vary in size from a typical psycholinguistic experiment to a large corpus, and in their repeated-measures structure. As we detail below, this lets us assess one of the known vulnerabilities of the bootstrap approach we describe later—its dependence on the source data.

Table 1. The number of per-word/region RTs and total number of subjects and sentence items in each of the three source data set used in our studies, depending on the data preparation applied (columns). The data preparation steps are applied cumulatively from left to right, as described in the next section.

pre-exclusion	post-exclusion	residualized	3-word region
HS18			
184572 RTs	183713 RTs	183713 RTs	147535 RTs
(mean=341, SD=694)	(mean=333, SD=162)	(mean=-6.20, SD=127)	(mean=-6.84, SD=92.3)
(min=1, max=226704)	(min=101, max=1999)	(min=-567, max=1769)	(min=-331, max=1274)
210 subjects	210 subjects	210 subjects	210 subjects
88 items	88 items	88 items	88 items
F13			
19000 RTs	18971 RTs	18971 RTs	15124 RTs
(mean=356, SD=150)	(mean=355, SD=139)	(mean=-10.18, SD=107)	(mean=-9.37, SD=77.4)
(min=45, max=5132)	(min=101, max=1977)	(min=-388, max=1590)	(min=-247, max=708)
38 subjects	38 subjects	38 subjects	38 subjects
51 items	51 items	51 items	51 items
NSC			
785102 RTs	763315 RTs	763315 RTs	688958 RTs
(mean=367, SD=5685)	(mean=333, SD=152)	(mean=-2.29, SD=127)	(mean=-2.08, SD=93.7)
(min=0, max=3944830)	(min=101, max=1998)	(min=-551, max=1755)	(min=-456, max=1680)
167 subjects	167 subjects	167 subjects	167 subjects
486 items	486 items	486 items	473 items

HS18. The first data set comes from a web-based SPR study by Harrington Stack and colleagues (Harrington Stack et al., 2018; henceforth HS18). A total of 481 subjects each read 144 sentences, one per

trial. After excluding subjects with low comprehension accuracy (<80%) and sentences with unintended mistakes, a total of 415 subjects and 124 trials remained. Here we use only the data from filler sentences. Since only one of the two between-subject conditions had filler sentences distributed across the entire experiment, we use only the data from that condition (the "Filler-first" group). This leaves RTs from 88 trials for each of 210 subjects (pre-exclusion, Table 1). As most SPR studies avoid placing critical regions in the sentence-initial/final position, RTs from the first and last words in each sentence are also excluded from HS18 and all other source data sets. Our simulations are based on the remaining 184,572 per-word RTs.

F13. Although web-based paradigms have been gaining popularity in psycholinguistic research, their poorer temporal precision and reduced levels of experimental control might impact the RT distribution. In order to examine similar data captured in a laboratory setting, our second source data set comes from an SPR study conducted in a university lab (Fine et al., 2013). The design of F13 is similar to that of HS18—in fact, HS18 was designed as a replication of F13 (for discussion, see Jaeger, Bushong, & Burchill, 2019). In Fine et al.'s study, a total of 80 subjects each read 71 sentences in the same basic paradigm as Harrington Stack et al. (2018). We again use only the RTs of fillers from the "Filler-first" condition. After excluding problematic subjects (for details, see Fine et al., 2013), and RTs from the first and last words of each sentence, our simulations are based on the remaining 19,500 per-word RTs from 51 trials for 38 subjects (see Table 1).

NSC. To further probe the generalizability of our findings, we choose our third set of data from a corpus study, rather than an experimental paradigm. The Natural Stories Corpus (NSC) is a corpus of reading times from ten stories edited to include more low-frequency syntactic constructions, while still sounding fluent to native speakers (Futrell et al., 2018). The original corpus was collected from 181 subjects, who each read five stories (each 33-64 sentences; chosen randomly), except for 19 subjects who read all ten stories. This source data thus has a repeated measures structure that differs from HS18 and F13, nesting sentences under stories. Following Futrell et al.'s comprehension-based exclusion criteria, 89 story instances are excluded. We further exclude 13 subjects with >75% of RTs smaller than 100 ms or larger than 2000 ms. Excluding

RTs from first and last words of each sentence, our simulations use the remaining 785,102 per-word RTs from 486 sentences and 167 subjects.

Distributional properties of RTs in the source data

Reading times, like reaction time data, exhibit soft lower bounds (for a demonstration, see SI-1.1; Luce, 1986) and long rightward tails, leading to high positive skew and kurtosis. This is also the case for all three of the source data sets, as illustrated in Figure 1a for HS18 (identical figures for the other two data sets are provided in SI-2.1). However, analyses of reading data are rarely based on raw uncorrected per-word RTs. We thus consider three data preparation steps that researchers might employ in a typical psycholinguistic study: 'outlier' exclusion or censoring, correcting RTs for word length, and/or aggregating per-word RTs into mean RTs over sentence regions (in order to reduce issues of word-to-word auto-correlations). Figure 1b-d shows that these common steps—cumulatively applied—do not remove the positive skew and kurtosis.

First, we apply a common criterion for outlier exclusion for SPR data by removing RTs ≤ 100 ms or ≥ 2000 ms from the analysis (Heider et al., 2014; Jegerski, 2014; Marsden et al., 2018; Nicklin & Plonsky, 2020). For HS18, this excluded < .5% of the data (see Table 1). This procedure partially reduces, but does not remove, the skew and kurtosis of the remaining RTs (Figure 1b). Second, we additionally apply standard corrections for word length effects and individual differences in reading speeds. This actually *increases* skew and kurtosis, compared to the previous step (Figure 1c).[1] Third, we apply aggregation of word RTs into region RTs, as is often done in SPR analyses. Specifically, we aggregate all word-by-word length-corrected outlier-removed RTs into three-word regions. This, too, does not remove the skew and kurtosis of those average per-word RTs (Figure 1d). Finally, Figure 1e shows that similar skew and kurtosis are also found for the critical three-word regions analyzed in the original study (Stack et al., 2018).

[1] Word-length corrected RTs are typically obtained by either fitting linear models with an intercept and word length effect separately to each subject's data, or by fitting an LMM with the same fixed effects plus random by-subject intercepts and slopes for the word length effect (for a critique of residualization, see Wurm & Fisicaro, 2014). Here, we follow the latter approach.

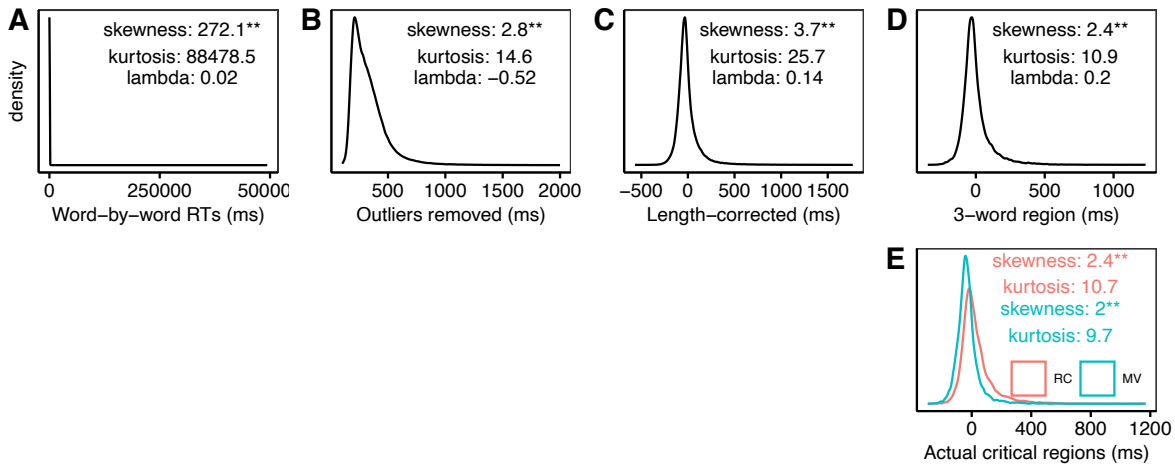

Figure 1. Marginal density of word-by-word reading times (RTs) using filler-trial data from Harrington Stack et al. (2018), Experiment 2. ***Panel A****: Raw RTs. Note that the skew is so large that the density appears as an 'L'.* ***Panel B****: Raw RTs after outlier exclusion as employed by Harrington Stack et al. (2018): removing RTs ≤ 100 ms or ≥ 2000 ms (< .5% data loss).* ***Panel C****: Length-corrected RTs after outlier exclusions, which also correct for individual differences in reading speeds between participants.* ***Panel D****: Average length-corrected RTs for a randomly chosen three-word region.* ***Panel E****: The length-corrected RTs averaged across the critical regions analyzed in the original study: relative clause (RC) and matrix verb (MV) sentences (Harrington Stack et al., 2018). Skewness is calculated as* $E[((X-\mu)/\sigma)^3]$ *and the kurtosis is the excess kurtosis calculated via Pearson's measure of kurtosis,* $E[((X-\mu)/\sigma)^4]$*. All distributions are significantly skewed based on D'Agostino test (D'Agostino, 1970). Box-Cox λs indicate which power transform would make the distribution of residuals most normal (see text for detail). Identical figures for the F13 and NSC source data are provided in SI-2.1.*

RTs also exhibit strong correlations between their means and standard deviations (Figure 2), violating the homogeneity of variance assumption. These correlations are not just due to individual differences in processing speeds between readers—correlations between RT means and standard deviations are evident within each subject (Figure 2a).

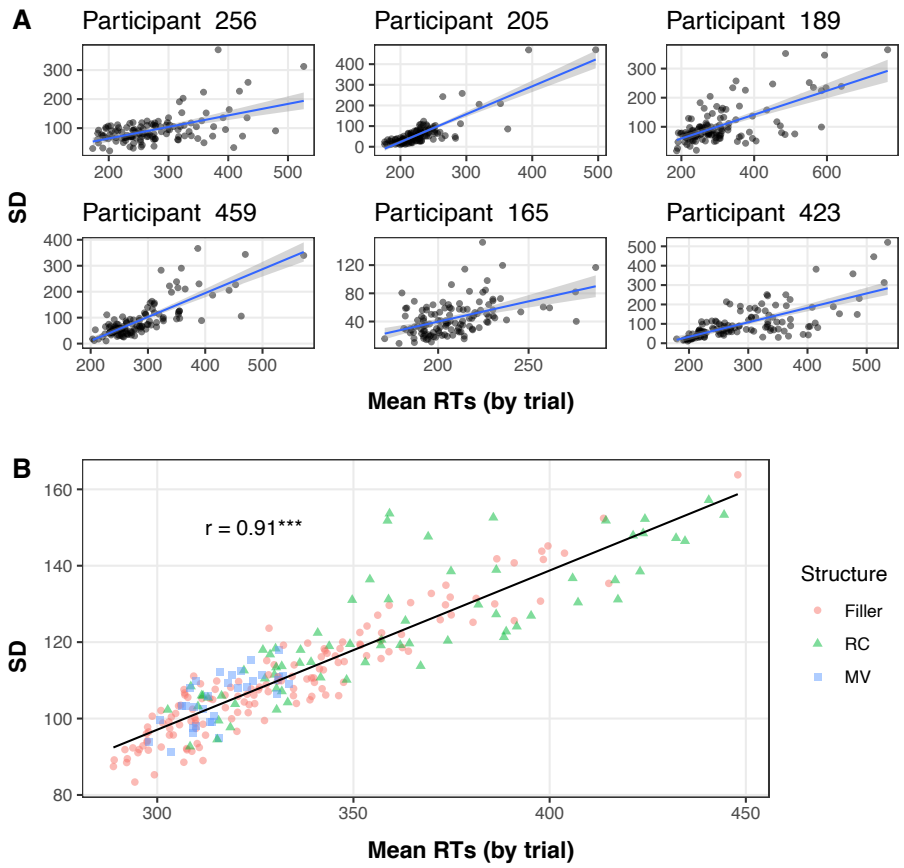

Figure 2. Correlation between means and standard deviations (SD) of outlier-excluded word-by-word reading times (RTs). ***Panel A****: Randomly drawn participants from Harrington Stack et al. (2018), Experiment 2. Points show mean per-word RTs of sentence items. Note that axis limits vary across panels.* ***Panel B****: Correlation across sentence items. Points show mean per-word RTs of sentences averaged across subjects. Shape and color show the sentence structure: filler sentences, relative clause (RC) sentences, and matrix verb (MV) sentences. RT means are a highly significant linear predictor of SDs both with and across sentence structures (ps<.0001). Identical figures for the F13 and NSC source data are provided in SI-2.1.*

Correlations between the mean and standard deviations are unexpected if RTs were normally distributed but expected if RTs follow, for example, log-normal distribution (Wagenmakers & Brown, 2007). To determine whether generalized linear models with alternative outcome distributions—such as the log-normal distribution—provide a better fit to their data, researchers often employ the Box-Cox test, which maximizes the log-likelihood of the residuals over a family of power transformations (Box & Cox, 1964; for a brief critique, see Sakia, 1992). The λ parameter returned by the Box-Cox test identifies the power transformation of the data that would make the distribution of residuals most Normal: e.g., λ=1 indicates the identity transform, λ=0 indicates the log-transform, and λ=-1 indicates the inverse or reciprocal

transform. Researchers then apply the chosen power transform to their data before using LMM or similar Normality-assuming approaches for analysis.

Applied to our three source data sets prior to exclusions, we find Box-Cox λs close to zero (HS18: 0.02, F13: 0.12, NSC: 0.06, with 95% CI widths <0.05), suggesting the log-transform for all three source data. This changes once exclusions are applied: after exclusions, and regardless of whether the data were additionally residualized and aggregated, the best-fitting λs for all three source data are close to -0.5 (see Figure 1b-d for HS18 and SI-2.1 for F13 and NSC). This would seem to suggest the rarely used inverse-square-root transform (i.e., $x^{-1/2}$). We emphasize that the specific λ values are of lesser importance here.[2] What is of potential relevance is that the best-fitting λs seem to vary substantially depending on the data preparation step, but show little variation across the three source data—despite the fact that the three source data were elicited in very different paradigms (web- vs. lab-based; factorial design vs. corpus). We return to this point below.

Generating RT data: parametric and non-parametric approaches

Having characterized the distributional properties of RTs in the source data, we describe two types of approaches to data generation—as employed in Type I error and power simulations—and ask how closely the data they generate approximate the distributional properties of natural RTs. The first type of approach follows previous work on reaction times in simpler psychometric paradigms, and generates RTs parametrically (Brysbaert & Stevens, 2018; Lachaud & Renaud, 2011; Liceralde & Gordon, 2022; Ratcliff, 1993; Schramm & Rouder, 2019). This includes, but is not limited to, the parametric assumptions corresponding to the most common power transformations used in RT analyses. Such parametric approaches are also the source of almost all power estimates provided in the reading literature. The second

[2] The outcome of the Box-Cox test depends on the relation between the actual distribution of RTs in the specific data set and the model for which the residuals are calculated. We used the R function *boxcox* from the *MASS* package (Venables & Ripley, 2002) to apply the Box-Cox test to an intercept-only linear model over all RTs in the source data, *lm(RT ~ 1)*. Other work has found, for example, the reciprocal transform to provide the best fit for RTs in the critical regions of other data sets (e.g., Nicenboim et al., 2014; Stone et al., 2020; Vasishth et al., 2013).

type of approach to data generation avoids parametric assumptions, and instead resamples RTs from the source data. We apply both types of approaches to all three source data, each under two different data preparation conditions: either prior to outlier exclusions or post exclusion of RTs ≤ 100 ms or ≥ 2000 ms (the two preparation steps that had the largest effect on Box-Cox λs for the source data).

Parametrically generated RTs. Following previous work (e.g., Brysbaert & Stevens, 2018; Lachaud & Renaud, 2011; Liceralde & Gordon, 2022), we fit separate LMMs with by-subject and -item random intercepts[3] to each of the three source data, and generate many new RT data sets from these LMMs. This approach has a number of advantages. It is conceptually transparent, and comparatively easy to program. The use of LMMs further makes it possible to capture by-subject and by-item variability, and even to sample new subjects and items.

To compare the consequences of different parametric assumptions, we fit the LMMs either to raw RTs or to transformed RTs. Specifically, we consider three Box-Cox transformations: the log-transform *log(RT)* and reciprocal transform *1/RT* (both of which are commonly employed in the reading literature), and the inverse square-root transform *1/sqrt(RT)* (which—while rarely employed in RT analyses—is suggested by the Box-Cox tests for all three of our post-exclusion source data sets). Inspired by a reviewer's suggestion (Marc Brysbaert), we also simulate a comparatively simple extension of the log-transform, the log-shift transformation, by first subtracting a lower bound from all RTs and then log-transforming them, *log(RT – lower bound)*. For each source data set, we set the lower bound to the minimum RT observed, minus 1 (to avoid NAs for the minimum RT). Unlike power transforms, the log-shift captures the fact that the soft lower bound for RTs is not zero but rather some positive value (Rouder, 2005; Schramm & Rouder, 2019). Adjusting RTs for this lower bound can be seen as a crude approximations of perceptual decision-making models, such as the drift diffusion model (Ratcliff, 1993; Ratcliff & Van Dongen, 2011). The log-shift transform thus provides a simple but potentially effective adjustment to the log-transform, and serves as an

[3] RT ~ 1 + (1 | SubjectID) + (1 | ItemID).

additional baseline against which we can compare the power transformations commonly applied in reading analyses.

For each of the four power transformations and the log-shift transformation, we then generate 10,000 simulated experiments from each LMM, each with 64 subjects and 64 items (here: words), while allowing for new subjects and items. Finally, we un-transform the generated transformed RTs back into raw RTs (e.g., for an LMM fit to log-transformed RTs, we exponentiate the generated log-RTs). This results in 5 (parametric assumptions) x 3 (source data) x 2 (pre- or post-exclusion) x 10,000 = 300,000 RT data sets, each containing 4,096 RT observations. For the post-exclusion data, we exclude RTs ≤100 ms or ≥2000 ms both prior to fitting the LMM and then again after RTs had been generated. This excludes approximately 7% and 1% of the generated identity transformed (raw) and reciprocal transformed RTs, respectively, and <0.3% for the other transformations.[4]

Bootstrapping natural RT distributions. As an alternative to parametric data generation, we employ a form of non-parametric hierarchical bootstrap. Instead of fitting a model and then generating data from it, we randomly sample RTs from the source data with replacement into bootstrapped data sets (BATAs). This approach avoids the strong parametric assumptions of the approach taken in previous work. It would, however, be wrong to consider it assumption-free. Instead, the non-parametric bootstrap makes a strong assumption: that the source data is representative of the population one is interested in understanding (e.g., the distribution of RTs in particular experiment, type of paradigm, or across paradigms).[5] This disadvantage of the bootstrap is ameliorated, the more source data one has from a given population and the more representative those data are (Efron & Tibshirani, 1994; Hinkley, 1994). This is the reason we employ source data sets of different sizes, repeated measure structure, and elicited from different paradigms. By

[4] While this second censoring step distorts the parametric distribution of the post-exclusion data, it guarantees that potential differences in RT distributions between the parametric and non-parametric bootstrap approaches are not trivially due to censoring (since bootstrap will never yield RTs outside the RT range found in the source data; see below). Additional analyses confirmed that the deficiencies of parametric approaches reported below showed even more clearly when censoring was applied only prior to data generation.

[5] A weaker form of this assumption is shared by parametric approaches, which assume that the parameter estimates obtained by fitting the model to the data are representative of the true population parameters.

comparing across different paradigms, we can begin to see whether BATAs from different paradigms have similar distributional properties. And, by comparing across source data sets of different size, we can begin to see whether the results of bootstrap analyses are reliable even for smaller data sets.

The specific approach we employ is a form of *hierarchical* bootstrap (for details, see SI-2.2). For each BATA, we first sample with replacement 64 sentence items from the source data. Within each of the sampled item instances, we sample a word position. Each of these instances of sentence plus word position is then assigned a unique item ID for the BATA (i.e., if the same combination of sentence plus word position was sampled twice, these two instances will have distinct item IDs in the BATA). Then we sample 64 subject instances in the same way we sample item instances, with replacement. The resulting BATA thus maintains inherent correlations between repeated measures of the same item across subjects, and repeated measures of the same subject across items. It also maintains inherent differences across subjects and items (for the NSC source data, for which subjects read stories, rather than isolated sentences, we additionally consider story as part of the hierarchical structure, see SI-2.2). Using this procedure, we sample 10,000 BATAs for each of the three source data under the two different data preparation conditions, resulting in an additional 60,000 RT data sets.

Distributional properties of parametrically and bootstrap generated RTs

To compare the RT distributions that result from the different approaches to data generation, we calculate the Box-Cox λ for each of the parametrically- or bootstrap-generated data sets. Before we present these results, we note that both the LMM fit to pre-exclusion raw RTs and the LMM fit to pre-exclusion inverse-square-root transformed RTs generated a large proportion of negative RTs (for raw RTs: 40-48% for the three source data sets; for inverse-square-root transformed RTs: 32-41%). On the one hand, the proportion of generated negative RTs will decrease with increasing RT mean and decreasing RT variance of the source data (e.g., the two approaches generated far fewer negative RTs when applied to the post-exclusion data). On the other hand, the fact that these two parametric approaches *can* generate negative RTs calls into question their suitability for reading time analyses (since the Box-Cox test is only defined for

positive input values, we excluded RTs ≤ 0). Similarly, the fact that the log-transformed or inverse-square-root transformed approach can generate RTs much smaller than the actual lower bound of natural RTs presents an *a priori* argument against those approaches (see also Schramm & Rouder, 2019).

Figure 3 shows the distribution of Box-Cox λs for the different data generation approaches, depending on whether RTs were generated from the pre-exclusion or post-exclusion source data (additionally analyses in SI-2.3 compare the data generation approaches in terms of correlations between the mean, variance, skew, and kurtosis of RTs). As would be expected, the hierarchical bootstrap yields λ distributions that vary around the λs of the source data. The range of λ values in these BATAs provides an indication of the range of λs one might obtain by replicating the source data sets by sampling new subjects and/or items. For all three source data, this range spans >0.5 units, suggesting that even exact replications of the an experiment might yield λs that suggest different power transforms.

A look across the three columns of Figure 3 further highlights how *little* the distribution of λs varies across BATAs from the three source data sets, despite the differences in the paradigms the RTs were elicited from. What does, however, strongly affect the distribution of λs is the censoring of the data from pre- to post-exclusion. Both of these simulation findings replicate—for both the non-parametric bootstrap and the parametric data generation approaches—what we found in the source data: a researcher's decision to censor ('outlier exclusion') can affect the distributional properties of RTs much more than differences in linguistic materials or subject populations. While this is unlikely to hold universally—e.g., the three source data we employ are comparatively homogenous in that they all reflect self-paced reading by 'typical' L1 speakers of a language spoken in a western, educated, industrialized, rich, and democratic ('WEIRD') culture over 'grammatical' (though sometimes infrequent) word sequences with well-formed (though sometimes implausible) meanings—it serves as a reminder that decisions in preparing reading data for analysis can strongly distort RT distributions, relative to the inherent variability in those distributions across data sets that arises naturally from between-subject and -item differences. In particular, our simulations suggest that outlier exclusion or data censoring might affect the distributional properties of RTs more strongly than some other common data preparation steps, like residualization or region aggregation.

Turning to the primary question of Study 1, parametrically-generated RTs resulting from any of the four power transforms we considered (identity, log, reciprocal, and inverse-square-root) do *not* match the λ distributions of bootstrapped natural RTs particularly well across the pre- and post-exclusion condition. Of note, RTs generated under the assumption made in the majority of RT analyses (normality) provide a consistently bad fit to the λ distributions observed in *any of* the three source data or the BATAs under *either* of the two data preparation steps (as do RTs generated under the reciprocal transform). RTs generated under the log-transform seems to match the λ distributions for the pre-exclusion BATAs, whereas data generated under the (rarely, if ever) used inverse-square-root transform seem to match the λ distributions for the post-exclusion BATAs. Additionally, both approaches match primarily the *mean* of the λ distributions of the BATAs, and fail to capture its spread and skew. For example, most parametric approaches predict less variability in λs than is actually observed when the data is bootstrapped.

Finally, the log-shift transformation matches at least the mean of the λ distributions of *both* the pre- and post-exclusion BATAs comparatively well. This makes sense: prior to exclusions—when the lower minimum RTs are 0 or close to 0 (see Table 1)—the log-shift approach generates essentially the same type of RT distribution as the log-transform (which *assumes* a lower bound of 0); however, once the data are censored to have a lower bound of 100 ms (post-exclusion), only the log-shift approach continues to be able to capture the log-normal-like nature of RT distributions because it avoids generating RTs that are never observed in the censored data ($RT \leq 100$). Under this interpretation, the seemingly good fit of the inverse-square-root transform for the post-exclusion data might be a coincidence—it simply reflects the fact that *within* the set of Box-Cox transformations, it is the closest approximation of a log-shift transformation with a lower bound of 100 ms. We suspect that this is not an uncommon scenario for the Box-Cox test when applied to the analysis of RTs: a power transform might appear to be a good fit for a particular data set simply because *none* of the power transforms capture the true generative structure of RT data.

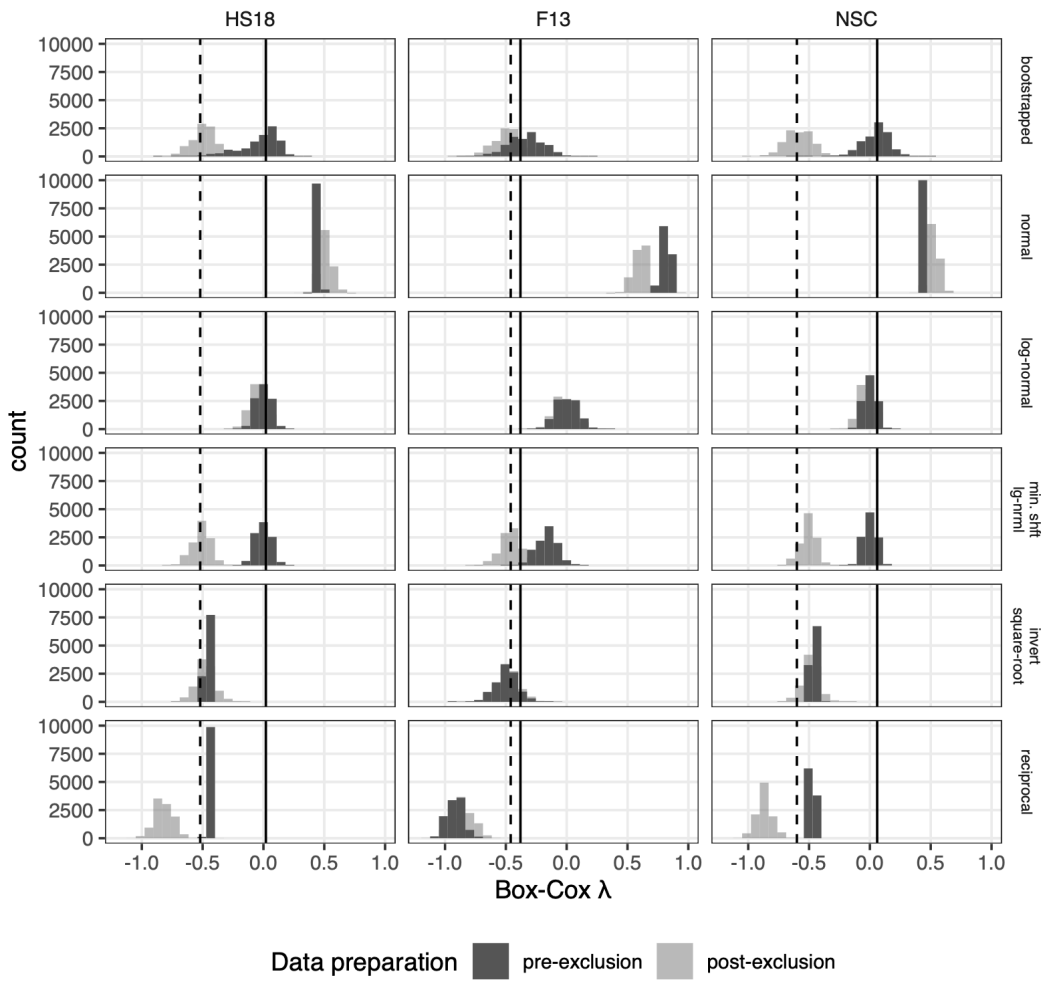

Figure 3. Distribution of Box-Cox λs of bootstrapped (top row) and parametrically-generated (remaining rows) RTs for all three source data sets. Each row shows a different data generation approach. Vertical lines indicate the Box-Cox λs of the source data pre-exclusion, dotted lines indicate the same post-exclusion (95% confidence intervals are too small to be visible in the plot).[6]

Discussion

Study 1 compared the distribution of natural RTs in three different source data sets that vary substantially in size (including number of subjects and items), stimulus structure (isolated sentences vs. connected stories), and the paradigms used to elicit them (e.g., lab- vs. web-based). We find that the

[6] Note that the distribution of λs is *not* always centered around the value expected for the transformation (e.g., 0 for the log-transform or -1 for the inverse-transform). This is a consequence of censoring RTs to be positive (for pre-exclusion data) or between 100 to 2000 msecs (for post-exclusion data), similar to what would happen during reading time analyses.

assumptions of *t*-tests, ANOVAs, and LMMs over raw RTs do not provide a good fit against the distributions of actual RTs, regardless of power-transforms (as assessed here via Box-Cox λs and confirmed through the analysis of statistical moments in SI-2.3).

Auxiliary analyses presented in SI-2.3 further show that bootstrapped natural RTs exhibit correlations between their mean, skew, and kurtosis that none of the parametrically-generated data sets exhibit. Additionally, some of these properties—such as the correlation between RT means and standard deviations found in all BATAs—are unexpected even if RTs were the *sum* of many different normal distributions, each with a different mean (due to, e.g., different lexical, syntactic, etc. properties of the words or context for which RTs are measured). This suggests that the mismatching λ distributions for normally generated RTs are not simply due to the specific model for which those λs were calculated in Study 1 (which did not model any word or contextual effects).

Of the power transforms commonly employed in reading research, only the log-transform leads to distributions that approximate those of bootstrapped natural RTs. However, even this match is limited to one of the two data preparation approaches. The results of Study 1 thus replicate for self-paced RTs previous findings for reaction times in simpler psychometric tasks (Baayen & Milin, 2010; Wagenmakers & Brown, 2007). Finally, we find that the log-shift transform provides a more consistent fit to human RTs than any of the power transforms we considered. This suggests that one downside of power transforms is their failure to account for lower bounds in the RTs (Schramm & Rouder, 2019). It also raises the possibility that a relatively simple adjustment to reading analyses—subtracting the minimum RT from all RTs in the data before log-transforming—might provide a viable analysis approach.

What does this mean for the field? The clear majority of reading analyses continue to assume normally or log-normally distributed RTs (corresponding to the use of raw or log-transformed RTs in *t*-tests, ANOVAs, and LMMs over raw RTs). A recent survey of the SPR literature suggests that up to 98% of RT analyses use statistical analyses that require normally distributed residuals, with 66% analyzing raw (untransformed) RTs, and approximately 16% analyzed log-transformed RTs (Nicklin & Plonsky, 2020). Our own Google Scholar searches conducted in September 2023 aligned well with these numbers. This

lends continued urgency to the question as to whether these analysis approaches inflate Type I errors and/or waste statistical power. The remainder of our studies thus assess the practical consequences of the normality and log-normality assumptions shared by most reading analyses.

Study 2

Study 2 begins to address this question for a simple by-2 between-group design (e.g., a comparison of a treatment condition against control). We use the hierarchical bootstrap from Study 1 to generate BATAs, and calculate Type I error rates and power of LMMs over raw and log-RTs. Given the promising performance of the log-shift transform in Study 1, we also include a simplified log-shift analysis as a third analysis approach, against which to compare the two more common approaches. We do so for all three source data sets from Study 1 under all four data preparation conditions depicted in Figure 1 (i.e., with and without outlier exclusions, with residualization, etc.). In Study 3, we compare the results obtained for hierarchical bootstrap against those obtained when RTs are parametrically generated.

The design of Study 2 is shown in Figure 4. For each combination of source data and data preparation, we compare the consequences of different sample sizes and effect sizes (including null effects in order to calculate Type I errors). In total, we analyze data from 1.44 million BATAs—10,000 each for the 3 x 4 x 4 x 3 simulation conditions in Figure 4. Each of these 1.44 million BATAs is submitted to all three analysis approaches. The resulting 4.32 million LMMs are used to calculate power and Type I error rates for the two analysis approaches.

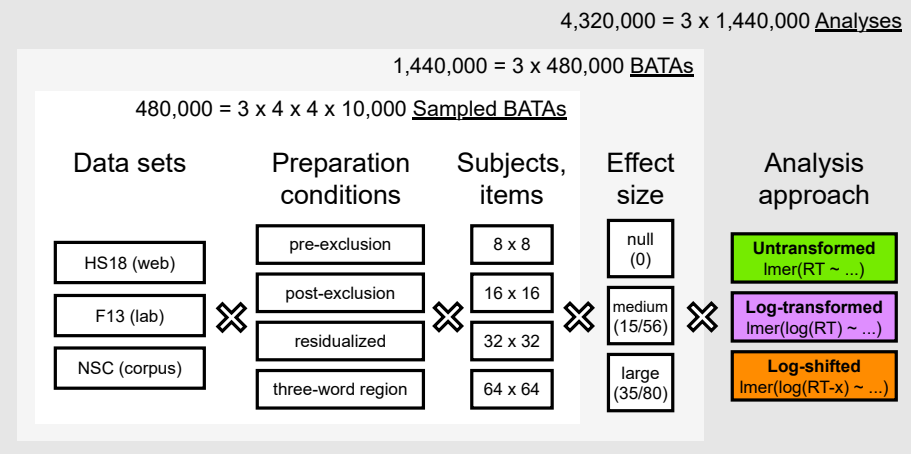

Figure 4. Simulation design of Study 2. Across the three data sets from which we bootstrap RTs, we draw on almost 1 million per-word RTs collected from over 600 sentences read by 400 subjects. Note that effect size labels ("medium" and "large") are to be interpreted relative to each other, and not to be confused with measures like Cohen's d. For details, see text.

Methods

The three data sets and four data preparation steps were described in Study 1. Next, we describe the two new simulation conditions and the analysis approaches shown in Figure 4.

Sample size. To cover common sample sizes of psycholinguistic experiments, we compare simulations with 8, 16, 32, and 64 subjects and items. Auxiliary Study 2a (SI-4) presents additional simulations that varied the number of items separately of the number of subjects. These simulations replicate all results of Study 2.

Effect size. Power simulations require the generation of *known* non-zero effects. For the HS18 and F13 BATAs, we simulate a between-subject design: we randomly assign half of the BATAs' subject IDs to Condition A and the other half to Condition B. We then subtract half of the simulated effect size from the RTs in Condition A and add half of the simulated effect size to RTs in Condition B (keeping the mean of the BATA's RTs constant across simulated effect sizes). To span a broader range of simulated designs, we instead simulate effects between *items* for the NSC BATAs (for details, see SI-3.1). For Type I error simulations, the procedure is the same except that we simulate null effects (i.e., we add and subtract 0).

Our goal is to *compare analysis approaches to each other*, rather than to obtain absolute power estimates. Specifically, we seek to determine whether one of the analysis approaches holds a *power*

advantage over the other, and whether any approach suffers from *inflated Type I error rates*. Since any potential power advantage—as well as which approach holds the advantage—might depend on the effect size, we consider two relative effect sizes. We use the labels "medium" and "large" for these two effect sizes, but emphasize that these labels have to be interpreted *relative to each other*, rather than as general measures of effect sizes (like Cohen's *d*). We first conducted pilot simulations to find a range of effect sizes for which power was neither at floor nor at ceiling. The effect sizes that met this criterion differed depending on whether the data had been residualized.[7] For the pre-residualization BATAs, we use 56 ms and 80 ms for the medium and large effect size, respectively. For post-residualization BATAs, we use 15 ms and 35 ms for the medium and large effect sizes, respectively. Finally, we generate BATAs with a 0 effect size for the Type I error rate simulations. The three different effect sizes (0 vs. medium vs. large) are always added to the exact same BATA, so that power and Type I errors are calculated over randomly sampled, but identical, BATAs.

All effects are added to BATAs' raw RTs—even when the analysis is conducted over log-transformed RTs. That is, BATA generation for Study 2 assumes that effects are linear in raw, rather than log-transformed, RTs. This is the assumption implicitly made whenever researchers use variants of the linear models over raw RTs. Study 2 thus generates effects in ways that *match* the assumptions of, and as such should favor, the untransformed approach. All data preparation conditions are applied *after* adding effects (for details, see SI-3.1).

Analysis approaches. We compare three types of analyses: LMMs over raw RTs, over log-RTs, or over log-shifted RTs, log(RT – lower bound). We note that LMMs over log-shifted RTS are less flexible than log-shift mixed-effects *regression models*. The latter allow researchers to *fit* the shift as a function of condition as well as by-subject and -item differences (e.g., Vasishth et al., 2018). This is a powerful

[7] An absolute effect of, e.g., 56 ms, will yield different power when applied to residualized RTs than when applied to pre-residualization RTs (since the two RT measures differ in variability, cf. Table 1). Comparisons of *absolute* power between pre-residualization and post-residualization would thus not be informative, regardless of our decision to add different effect sizes to pre-residualization and post-residualization data.

approach that can, however, come with additional convergence issues. It also requires *substantially* more compute time. While this is not typically an issue for a single analysis, it is a consideration for the millions of model fits required for Study 2.

We implement all analyses using the *lmer* function from the *lme4* package (Bates, Mächler, et al., 2015). For the pre-residualization BATAs, the analysis contained by-subject and by-item random intercepts.[8] For the post-residualization BATAs, only by-item intercepts are included, as the residualization already removes individual differences between subjects (and, indeed, additional simulations found that inclusion of by-subject intercepts for the post-residualization BATAs resulted in high rates of convergence failures). Distributed over 180 cores on a total of 30 machines, the BATA generation and analyses for Study 2 took about 5 days of compute time (for implementation details, see SI-3.2).

Results

Measuring the success of analysis approaches. For each type of simulation, we compare the Type I errors and power (1 – Type II error rate) of the two analysis approaches. Analyses that failed to converge were excluded from the calculation of Type I errors and power.[9] We define the Type I error as the proportion of BATAs for which the effect reached significance ($p<.05$) when the simulated effect was, in fact, null. Similarly, we define power as the proportion of BATAs for which a non-zero effect reached significance ($p<.05$) and went in the correct direction.

[8] Since our BATAs simulate between-subject (HS18 and F13) and between-item designs (NSC), the maximal random effect structure justified by the design would include random by-item slopes for the between-subject design and random by-subject slopes for the between-item designs. Initial simulations that included those slopes confirmed the results we report, but also suffered high rates of non-convergence (a general issue with maximal random effect structures, see Matuschek et al., 2017). While non-convergence for maximal random effect structures is not uncommon for real experiments either (Bates, Kliegl, et al., 2015), it is possible that non-convergence for our BATAs was caused by the fact that our bootstrap approach does not add parametrically-generated by-subject and -item variability *to the effect.* It does, however, inherit natural by-subject and -item variability in the intercept (which is typically the largest source of variability by far). Auxiliary Study 2e (SI-8) presents additional simulations in which we added parametrically generated by-item (HS18 and F13) or by-subject (NSC) variability to the *effect*. These simulations confirm the results reported here.

[9] Rates of convergence failures (always < 0.3%) and singular fits (up to 40% for the smallest BATAs) are reported in SI-3. The log-transformed and log-shift transformed approaches consistently outperformed the standard approach.

One concern for power simulations is that power can appear high because the analysis is anti-conservative, yielding significant effects at a higher than targeted rate even when there is no real effect. We thus report Type I error-corrected power. For each simulation condition, we first determine the significance threshold α' for which the Type I error rate would have been exactly $\alpha = .05$. We then calculate the Type I error-corrected power for that simulation condition as the proportion of BATAs that reach significance under the revised criterion ($p<\alpha'$) when the simulated effect was *not* null (i.e., for "medium" and "large" effects). Uncorrected power is reported in SI-3.6, and replicates all patterns reported in the main text.

Type I errors. Figure 5 summarizes the Type I error rates for all simulation conditions. To compare the degree of conservativity between the three analysis approaches, we conducted binomial tests comparing the Type I error rates in each simulation condition against the targeted rate of 0.05 (see shape of points in Figure 5). Only two simulation conditions result in anti-conservativity: the 8x8 and 64x64 pre-exclusion F13 simulations with the log-transformed analysis approach. Most analyses had Type I error rates significantly *lower* than 0.05 (27 out of 48 conditions for the untransformed analysis approach, and 28 for the log-transformed approach, and 33 for the log-shift approach). Additional analyses reported in SI-3.5 compare the Type I error rates of the untransformed and log-transformed approach, finding the latter to be less conservative and closer to the targeted Type I error rate. These results suggest that (1) all analysis approaches tend towards conservativity rather than anti-conservativity, and (2) the log-transformed approach most closely approximates the targeted Type I error rate. We discuss the implications of these results after presenting the power results.

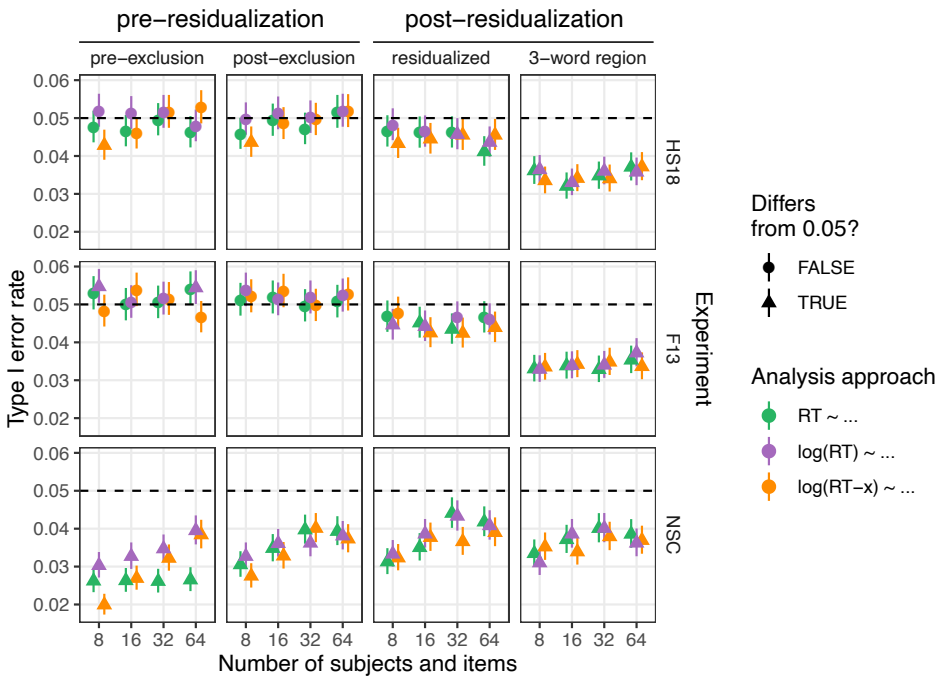

Figure 5. Type I error rate for all simulation conditions in Study 2. Type I error rates are generally close to the expected rate of 0.05 (dashed horizontal line). Significant differences from .05 (based on a binomial test) are marked by transparency. Error bars show 95% binomial confidence intervals.

Power. Figure 6 shows the Type I error-corrected statistical power for all simulation conditions for HS18. As would be expected, power increases with increasing sample size and effect size. The same pattern holds for the other two data sets (Figure 7).

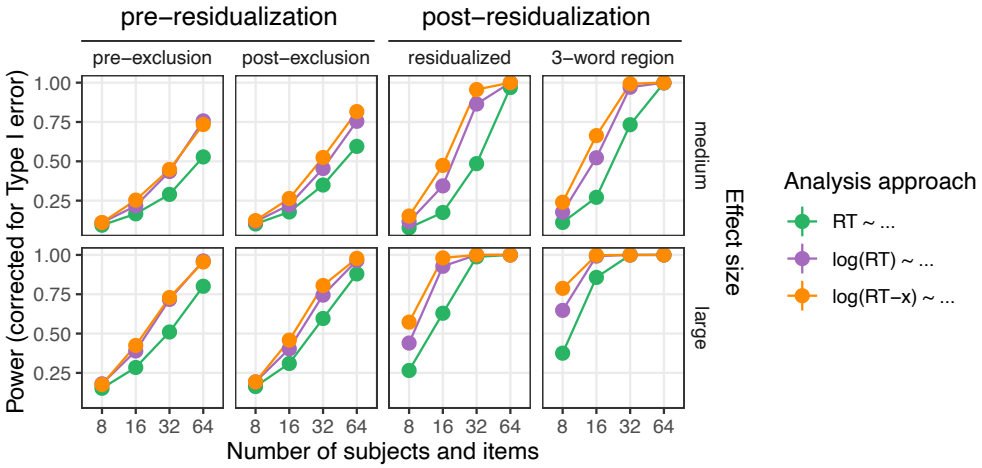

Figure 6. Power (corrected for Type I error rates) for all simulation conditions for the HS18 data in Study 2. Confidence intervals are shown but too small to be visible. See SI-3 for identical figures for the F13 and NSC data.

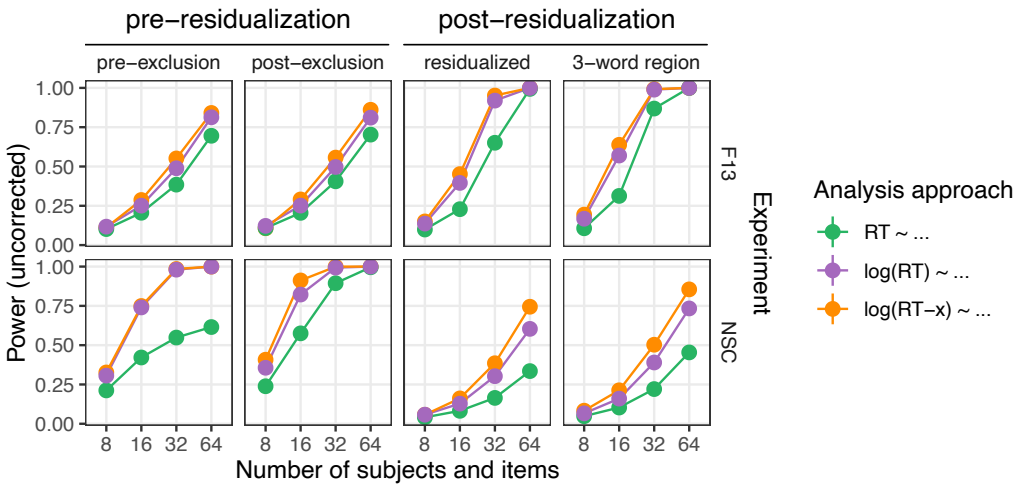

Figure 7. Power (corrected for Type I error rates) for the F13 and NSC data, averaged across effect sizes in Study 2. Confidence intervals are shown but too small to be visible.

Critically, Figures 6 and 7 also suggest that the log-transformed analysis approach has a power advantage over the untransformed approach. This was confirmed by logistic regression analyses reported in SI-3.7: for BATAs from all three source data, analyses over log-transformed RTs yields significantly higher power than analyses over untransformed RTs (*p*s<0.001). This power advantage is up to 2-3 orders of magnitude larger, and more consistent across both source data and simulation conditions than the small differences in Type I error described above: the log-transformed analysis approach yields higher power for *all* simulation conditions (except when both approaches had 100% power). And, the more evidence there is for an effect—i.e., with increasing sample size and effect size—the more pronounced is the power advantage of the log-transformed analysis (see Figure 8).

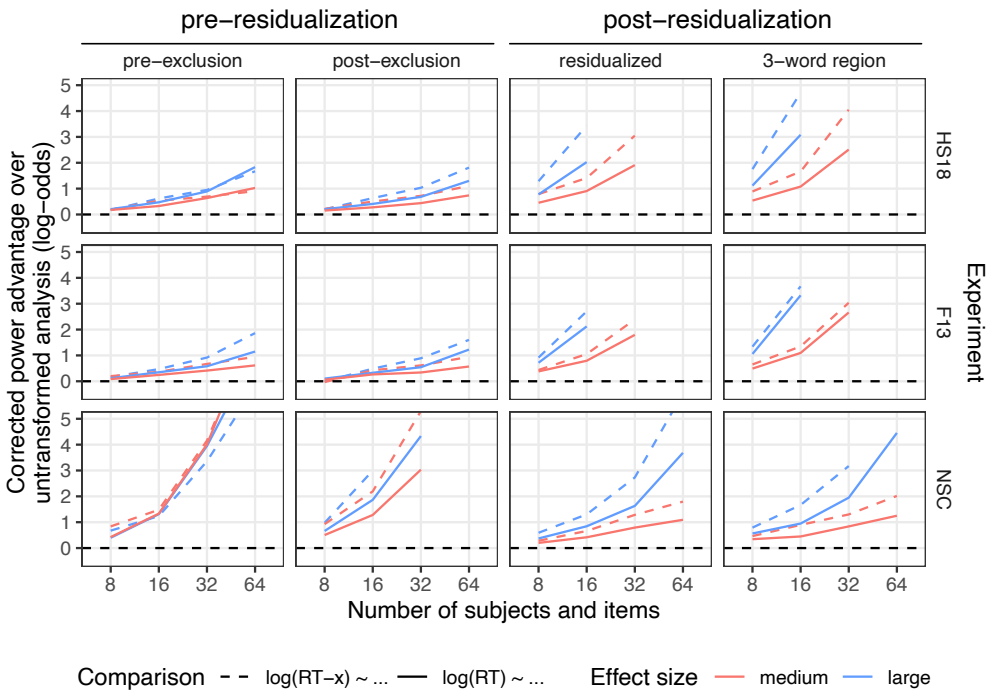

Figure 8. Relative power advantage (Type I error-corrected) of the log-transformed analysis approach in Study 2. Power advantage as compared to the untransformed analysis approach (as a difference in log-odds) for all three data sets and each effect size. Simulations for which power was at ceiling in either analysis approach are excluded (because they would result in infinite log-odds).

Finally, the log-shift transformed approach yields even higher power than the log-transformed approach (with the exception of the pre-exclusion condition, for which the two approaches yielded similar power). Replicating the pattern for the log-transformed approach, the power advantage of the log-shift approach increases with sample size.

Discussion

In every single comparison, for all three source data, all data preparation steps, and all simulated sample and effect sizes, LMMs over raw RTs are less likely to detect real effects than LMMs over log-transformed or log-shift transformed RTs. This power advantage of the two log-transformed approaches suggests that the ways RT data violate the analysis assumptions of LMMs can indeed have practical consequences. The fact that these results hold for all three source data provides some evidence that this power deficit is not specific to any single SPR experiment or paradigm. For any experiments for which power is not severely underpowered to begin with, we find that the power advantage of the two log-transformed approaches is

substantial: for the 16x16 simulations from the HS18 data for residualized RTs with a large effect size, the untransformed analysis approach has 62.9% power, and the log-transformed approach has 99.2% power—i.e., the odds of detecting an effect are 73 times greater in the log-transformed approach. Averaging over all other simulation conditions, the power advantage of the log-transformed approach over the untransformed approach ranges from 1.01 log-odds for F13 to 1.95 log-odds for NSC (1.32 for HS18). The average power advantage of the log-shift approach is even larger, ranging from 1.25 log-odds for F13 to 2.46 log-odds for NSC (1.67 for HS18). The magnitude of this advantage becomes apparent if we imagine an experiment with 80% power for the log-shift transformed analysis: in this case, LMMs over untransformed RTs would achieve only 53%, 25%, and 43% power, respectively (i.e., between 27-55% lower power)! We return to this issue in the general discussion but emphasize here that the specific power advantage of log-transformed and log-shift approaches is likely to vary between data sets.

Critically, the power advantages of log-transformed or log-shift transformed analyses do not seem to come at the cost of inflated Type I errors—at least not for the design considered in Study 2. Finally, additional analyses reported in SI-3 find that the log-transformed approach consistently outperformed the untransformed approach in terms of convergence and avoidance of (detected) singular fits. In short, for the experiment designs considered in Study 2, the log-shift transformed approach only seems to have advantages with little to no downsides. Of particular note is that this advantage is observed despite the fact that the BATAs in Study 2 generated effects to be linear in raw RTs—i.e., the way we added effects followed the assumptions of the *untransformed* approach, and yet we observed an advantage of the log-transformed approach (Auxiliary Studies 2d and 2e in SI-7 and SI-8 confirm the results of Study 2 when effects are instead generated to be linear in log-RTs).

Comparison to previous work. The power advantage of the log-transformed approaches stands in contrast to some previous investigations of reaction time data from simpler psychometric tasks (Liceralde & Gordon, 2022; Schramm & Rouder, 2019). These studies found that the untransformed approach had higher statistical power compared to analyses of power-transformed data (including log-transformed reaction times). We see two possible reasons for the difference in findings.

First, Study 2 analyzed reading times while previous work analyzes reaction times. Although reading and reaction times both constitute "chronometric data" and their distributions exhibit similar properties (e.g., right-skew, extreme outliers, lower bounds), the distribution of reaction time data is expected to depend on the specific tasks and the perceptual, cognitive, and motor mechanisms it requires (Ratcliff, 1993; Wagenmakers & Brown, 2007). The mechanisms underlying reading are likely different than those underlying simpler reaction time tasks.

Second, the BATAs analyzed in Study 2 are obtained by bootstrapping natural RT distributions. This contrasts with previous studies, all of which compare analysis approaches to *parametrically* generated data. As we have shown in Study 1, such parametric assumptions often do not provide a good fit against actual RT distributions. It is thus possible that the results of previous work were affected by the parametric assumptions made in those studies. Study 3 addresses this possibility by repeating the analyses of Study 2 over parametrically generated data.

Study 3

To illustrate the consequences of parametric assumptions for data generation, we generate RTs under two different assumptions—RTs with normal vs. log-normally distributed residuals—while otherwise exactly following the design and approach of Study 2. This covers two common parametric approaches to power simulations for reading studies. Beyond the immediate goal of understanding why our results differ from those obtained in previous research on other types of chronometric data, Study 3 also speaks to the interpretation of power estimates provided in reading research, for which parametric approaches remain the norm (e.g., using the popular *simr* package in R, Green & MacLeod, 2016). The design of Study 3 is shown in Figure 9, and described next.

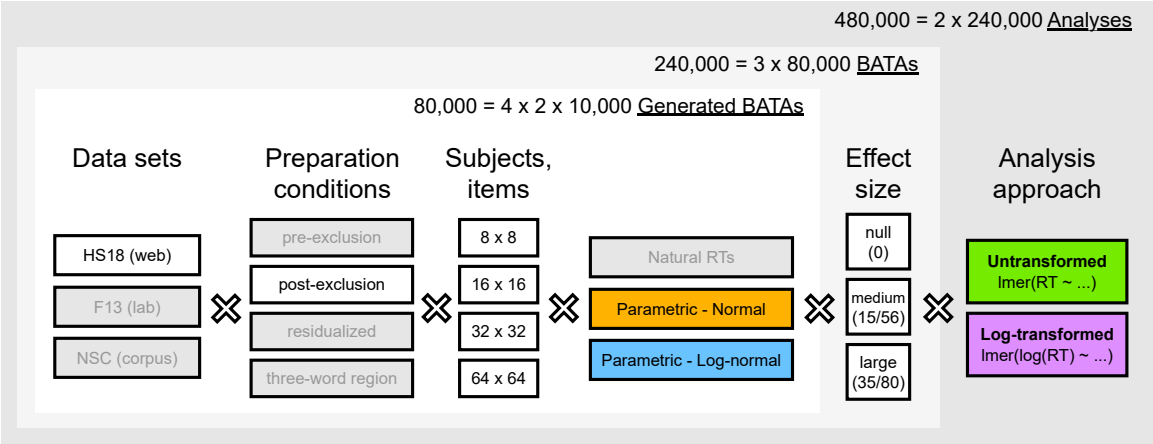

Figure 9. Simulation design of Study 3. Simulation conditions present in Study 2, but not included in Study 3, are grayed out.

Methods

Data and data preparation conditions. For the sake of computational feasibility, we focus on a single source data set (HS18) and data preparation step (post-exclusion). We choose HS18 because the majority of analyses in psycholinguistics continue to be based on factorial designs rather than corpus data, and HS18 is more than eight times larger than F13.

Sample sizes, effect sizes, and analysis approaches. We use the same approach as in Study 1 to parametrically generate data under two different distributional assumptions: normality and log-normality. For the normally generated RTs, we fit an LMM with a fixed effect intercept as well as random by-subject and by-item intercepts to the source data, and then generated RTs from the LMM. Finally, we add effects following the same procedure as in Study 2. For log-normally generated RTs, we follow the same process but fit the LMM to log-RTs, and then transform the generated log-RTs back into raw RTs before adding effects. The simulated effect sizes, sample sizes, and the analysis approaches were identical to Study 2.

Results

Type I errors. The Type I error rates of both sets of parametrically generated RTs closely approximated the expected value of 0.05, without significant conservativity or anti-conservativity (Figure 10). Nested model comparisons did not find any significant effects of analysis approach (see SI, SI-9.3). These results replicate Study 2 for the post-exclusion HS18 data (see middle row, post-exclusion panel of Figure 5).

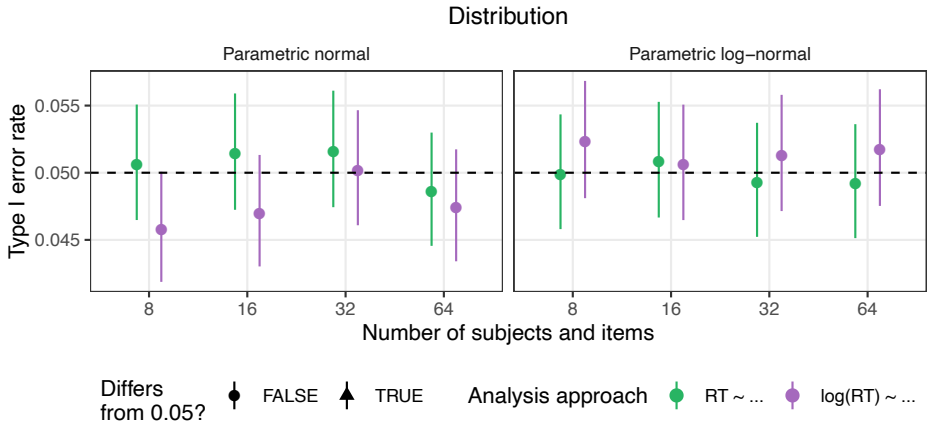

Figure 10. Type I error rates of both analysis approaches for parametrically generated RTs in Study 3. For additional details, see Figure 5.

Power. For normally generated RTs, the log-transformed approach had *less* power than the untransformed approach for every simulation condition that did not reach ceiling (Figure 11)—a reversal of the findings for bootstrapped RTs in Study 2. This power *dis*advantage of the log-transformed approach was significant (see SI-9.5), though smaller in magnitude (on average 0.15 log-odds) than the advantage for bootstrapped BATAs in Study 2 (0.524 log-odds for post-exclusion HS18 simulations).

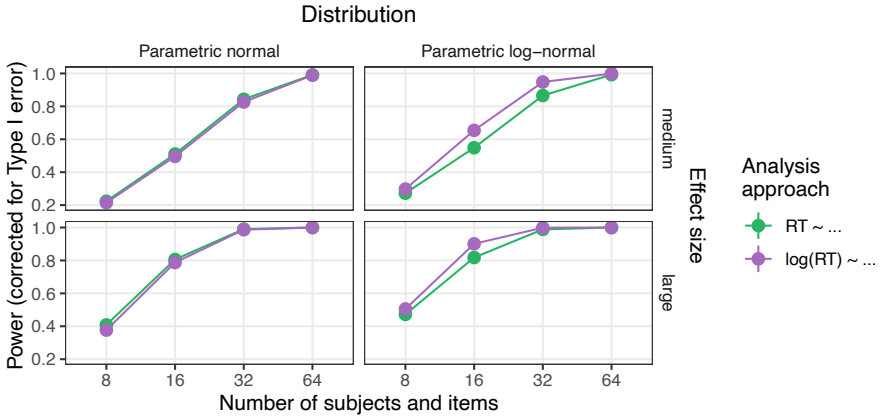

Figure 11. Power (corrected for Type I error rates) of both analysis approaches for parametrically generated RTs in Study 3. Normally generated RTs show the opposite *pattern as Study 1, with the untransformed analysis approach having greater corrected power for every simulation not at ceiling. Log-normally generated RTs, however, show the same pattern as in Study 1, with the log-transformed analysis approach having greater power.*

For log-normally generated RTs, on the other hand, we qualitatively replicate the results of Study 2: the log-transformed analysis approach had *more* power for every simulation not at ceiling (Figure 11). This

power advantage was significant (see SI-7) and *larger* (0.962 log-odds) than for the bootstrapped RTs in Study 2 (0.524 log-odds).

Discussion

The comparison of Studies 2 and 3 demonstrates that power estimates can *strongly* depend on the assumptions baked into the data generation. This is further summarized in Figure 12, which highlights two important consequences of parametric data generation approaches that do not match the natural statistics of RTs.

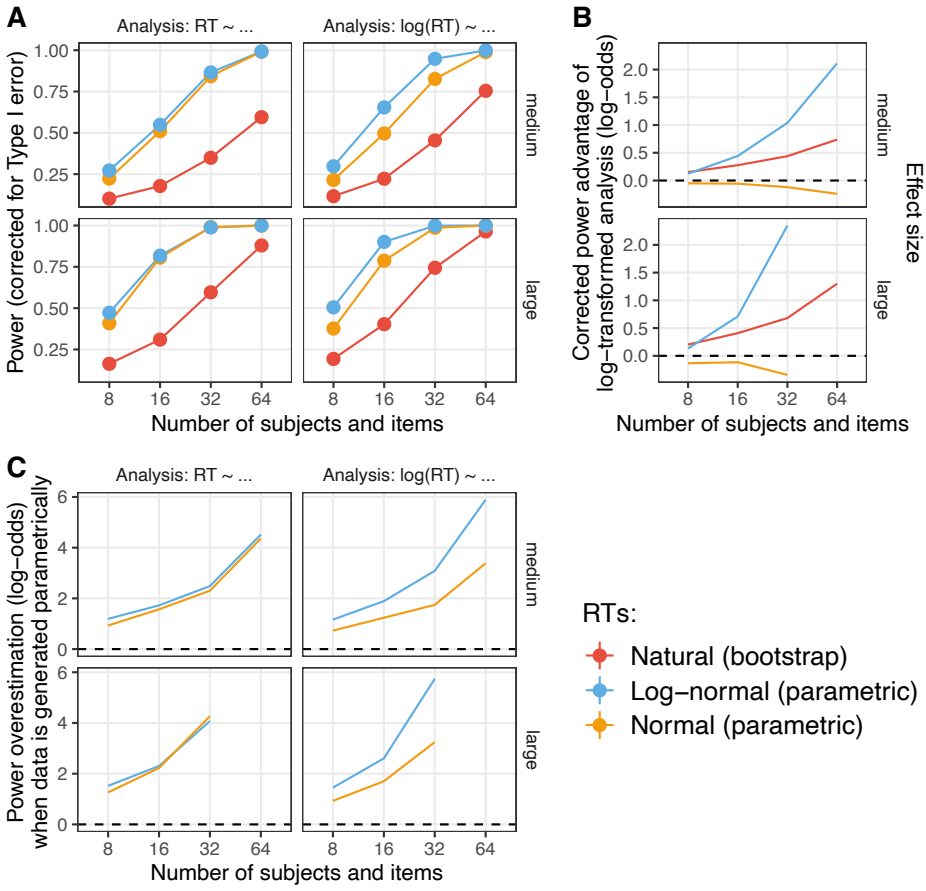

Figure 12. Comparing the parametric power simulations from Study 3 with the bootstrapped power simulations over natural RTs from Study 2. ***Panel A****: Type I error-corrected power. Note how the parametric power estimations group together and are much larger than power estimates based on bootstrapped RTs.* ***Panel B****: Type I error-corrected power* advantage *of log-transformed analysis over the untransformed analysis (as a difference in log-odds) for simulation conditions not at ceiling. Unsurprisingly, the analysis matching the parametric assumptions under which the data was generated yielded the most power. The power advantage for actual RTs falls between that for normally and log-normally generated RTs.* ***Panel C****: Degree of overestimation of Type I error-corrected power when the data*

is generated parametrically (as a difference in log-odds from Type I error-corrected power for the bootstrapped RTs). Notice that overestimation increases with sample size.

First, Figure 12a compares the consequence of different data generation approaches within analysis approach. Compared to the bootstrapped RTs in Study 2, both parametric data generation processes that we considered systematically *over*estimate the statistical power of both LMMs over raw RTs and LMMs over log-RTs (recall that both studies added the same effects and did so always in raw RTs). And, at least for the design considered here, degree of over-estimation *increases* for larger sample sizes (Figure 12c), presumably because the over-optimism of parametric power estimates increases with estimated power. This does not mean that *all* parametric data generation approaches would suffer from the same problem. For example, it is possible that more complex parametric approaches—such as log-shift models—would yield more realistic results. It does, however, highlight that two commonly employed parametric approaches to power simulations (assuming normality or log-normality) can lead to overly optimistic power expectations. This over-optimism can be substantial: in some cases, the parametric power simulations in Study 3 yield power estimates that are 50% higher than those based on bootstrapped natural RTs. For example, reading researchers using a power estimate based on the assumption of normality might mistakenly believe their experiment to have more than 95% power, when in reality their sample yields less than 50% power. Such over-optimism is arguably of particular relevance with the field's renewed focus on replication: the noteworthiness of replication failures depends on the *actual* statistical power of the replication, not the estimated power (see also Jaeger et al., 2019).

The second difference between parametric and bootstrap data generation approaches pertains to the conclusions they support about the suitability of different analysis approaches. Using bootstrapped natural RTs, Study 2 found that the log-transformed analysis approach had significantly more power than the untransformed approach. Study 3 demonstrates that one would come to a similar conclusion if the data had been generated under the assumption of log-normality, but to the opposite conclusion if the data had been generated under the assumption of normality. Figure 12b shows that the power advantage obtained from the non-parametric bootstrap approach from Study 2 falls in the middle of the estimates obtained from the

two parametric approaches employed in Study 3. These results suggest that previous comparisons of analysis approaches might have been affected by the assumptions made in their data generation approach. Previous work that has compared different analysis approaches for related psychometric data has largely relied on parametrically generated data (Brysbaert & Stevens, 2018; Lachaud & Renaud, 2011; Liceralde & Gordon, 2022; Ratcliff & Smith, 2004; Schramm & Rouder, 2019). Even when more complex—and perhaps more appropriate—parametric models are used to generate the data (Liceralde & Gordon, 2022; Rouder, 2005), there is always a risk that the parametric assumptions fail to capture important aspects of actual distributions of natural RTs (Palmer et al., 2011).

The hierarchical bootstrap procedure offers an alternative to parametric power estimates. While it is not without its own potential pitfalls—in particular, in its increased dependence on the assumption that the source data is sufficiently large and representative of the targeted population—the hierarchical bootstrap can complement parametric approaches. This can help avoid over-optimistic power estimates, and more reliably guide the choice of analysis approaches. We return to this point in the general discussion.

Interim Summary

What does all of this mean for the two approaches to RT analyses we have compared? In Study 1, we found that none of the Box-Cox transforms that are commonly applied to RT data (identity, log, or inverse transform) result in RT distributions that closely match the distributional properties of natural RTs. Critically, the results of Study 2 suggest that this has consequences for the reliability of different analysis approaches: LMMs over log-RTs reliably yielded more—often substantially more—statistical power, without causing inflated Type I error rates. This advantage generalized across three source data sets that employed different paradigms to collect RTs, and across different data preparation procedures, sample sizes, and simulated effect sizes. Five auxiliary Studies 2a-e presented in the SI further assessed various decisions we made in the generation of BATAs for Study 2. All of these auxiliary studies replicate the power advantage for log-RT analyses found in Study 2, and the lack of downsides in terms of Type I errors,

convergence, or singular fits. Finally, Study 3 suggests that conflicting results from previous research might be at least in part due to the parametric approaches to data generation used in those works.

Based on these results, one might be tempted to recommend that reading researchers stop using LMMs over raw RTs, and instead use LMMs over log-RTs (or other methods after appropriate validation). This would be good news for experimenters, as it suggests an easy 'fix' to *t*-tests, ANOVAs, and LMM analyses. However, the simulations in Study 2 as well as all auxiliary studies in the SI have an important limitation. All simulations we have presented so far investigated the main effect of simple by-2 designs (between-subjects or -items for, e.g., Study 2, within-subjects and -items for Study 2e). Critically, there are *a priori* reasons to expect different findings for factorial designs or other designs with *interactions*. Before we can make general analysis recommendations, Study 4 thus compares the two analysis approaches for a 2x2 factorial design.

Study 4

Study 4 expands Study 2 in two ways. Our primary goal is to illustrate how the analysis of *factorial* designs is affected by the choice of analysis approach. To this end, we simulate a 2x2 design with both main effects and interactions. As we discuss below, interactions are known to be particularly vulnerable to assumption about the scale in which an effect is assumed to be linear. We thus revisit an assumption made in Study 2: that the effects of interest are linear in *raw* RTs. Just like the normality assumption, assumptions about the scale are made in the majority of reading time analyses. This is true even for the more advanced analyses we mentioned in the introduction, which either *assume* linearity in one or more scales (e.g., shifted log-RTs for the log-shift model; exponentiated RTs or raw RTs in the exGaussian mixture model), or *penalize* against non-linearity in specific scales (e.g., generalized additive models, Wood, 2011). While some of these approaches can substantially ameliorate the strength of scale-dependence (e.g., generalized additive models), they do not completely free researchers from effects of scale assumptions (we return to this issue in the general discussion). The way we generated effects in Study 2 *matched* the assumptions of

linear analyses over raw RTs. Study 4 instead includes both simulations that generate effects linear in raw RT, and simulations that generate effects linear in log-RTs (see also Auxiliary Studies 2d and 2e in the SI).

The design of Study 4 is summarized in Figure 13. Unlike in Study 2, we do not consider analyses over log-shifted RTs. For reasons that we return to in the discussion, LMMs over log-shifted RTs are expected to exhibit qualitatively similar scale-dependence to LMMs over log-RTs. We start by describing how we added effect in raw or log-RTs. This provides the background necessary to appreciate how interactions are particularly vulnerable to assumptions about the scale in which effects are linear.

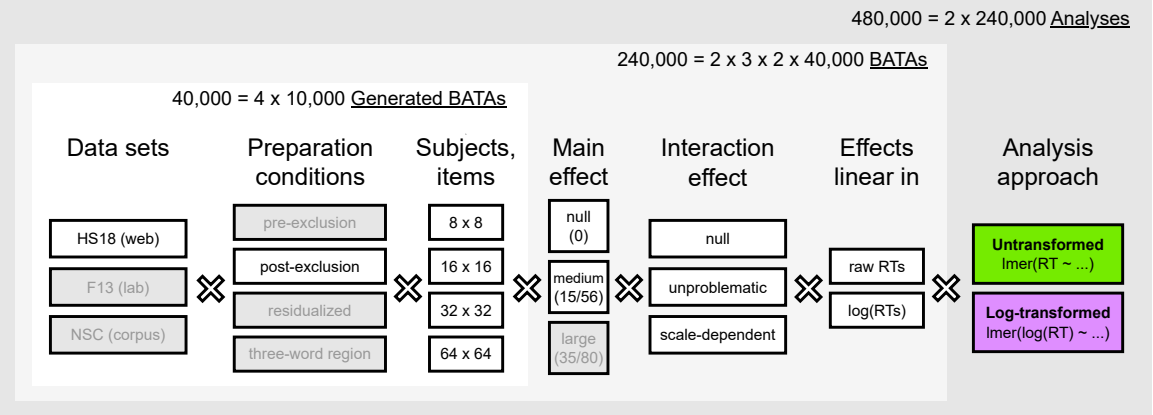

Figure 13. Simulation design of Study 4.

What scale are effects linear in? For Studies 2 and 3, we assumed that the effects that researchers try to detect are linear in raw RTs: i.e., if an effect adds 56 ms to an RT of 200 ms, it would add the same 56 ms to an RT of 1000 ms. While this is the assumption made in most RT analyses, it can be problematic. RT distributions are known to have soft lower bounds, reflecting the limits of human reaction times (Luce, 1986). In the presence of such bounds, effects can*not* be linear across arbitrary ranges of raw RTs.

This leaves open the possibility that the effects of interest to researchers are approximately linear in raw RTs *within RT ranges sufficiently larger than the lower bound*. There are, however, reasons not to take linearity in raw RTs for granted even in this more constrained sense. For example, just as some theories predict certain effects to be linear in raw RTs (Smith & Levy, 2013), other theories predict other effects to be linear in, e.g., log-transformed or log-shifted RTs (i.e. multiplicative for raw reaction times, e.g., for

memory cues, Van Dyke & McElree, 2006; and some models of reading, Lewis & Vasishth, 2005) or inverse RTs (e.g., as additive changes in processing rate, Carpenter & Williams, 1995). Even in the absence of clear theoretical motivations, linearity in raw RTs is not necessarily more plausible of an assumption than linearity in other scales (e.g., Kliegl et al., 2010). While some have appealed to ease of interpretation as an argument for linearity in raw RTs (e.g., Lo & Andrews, 2015; Osborne, 2002), ease of interpretation is hardly of value if it leads to wrong interpretations (because of unjustified assumptions about the scale in which effects are *actually* linear). To illustrate the effect of linearity assumptions, Study 4 manipulates whether effects are generated to be linear in raw RTs or log-RTs.

Scale-dependence of interactions. One aspect of analyses that depends particularly strongly on linearity assumptions are interactions. Previous research has distinguished between "scale-dependent" or "scale-independent" interactions (Liceralde & Gordon, 2022; Loftus, 1978; Wagenmakers et al., 2012). For example, effects that are actually additive in log-RTs are multiplicative in raw RTs, creating the illusion of a *super-additive* interaction in raw RTs (Figure 14, left panel). Conversely, an actual super-additive interaction in raw RTs can become undetectable when RTs are log-transformed (same panel).[10] Such interactions are considered "scale-dependent" (i.e., their existence is dependent on what scale/transformation of the data). Only "scale-*in*dependent" interactions—such as "cross-over" interactions, which are comparatively rare—persist regardless of the scale (Figure 14, right panel; Loftus, 1974). While these issues have received a fair amount of attention, including for reaction and reading time analyses (e.g., Lo & Andrews, 2015; Staub, 2020), it remains common practice to draw strong conclusions even from scale-dependent interactions.

[10] It is important to emphasize that one cannot conclude that an interaction that disappears after log-transforming RTs is necessarily not real, or that log-transforming RTs is wrong because it can remove interactions. What follows from an interaction (dis)appearing under different transforms depends on which scale the actual effects are linear in.

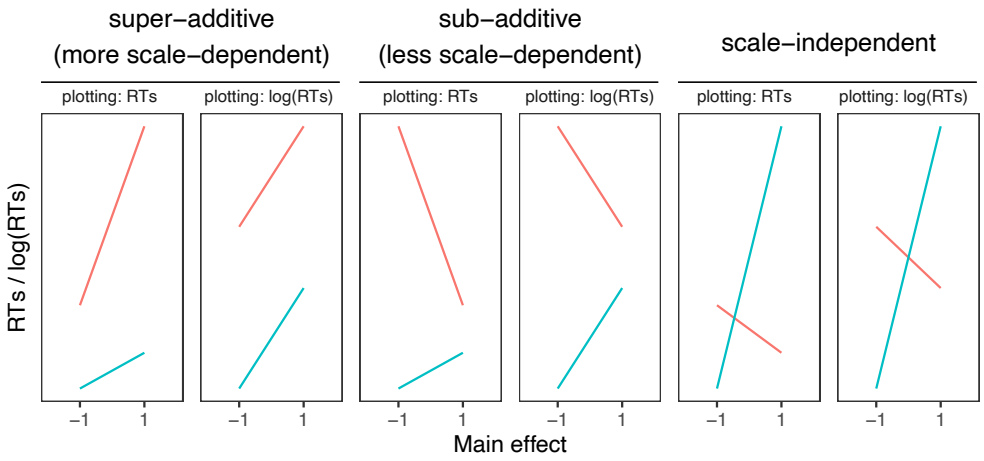

Figure 14. Illustrating a continuum of scale-dependence for interactions. ***Left****: scale-dependent interaction in raw RTs: the two effects have a super-additive relationship. However, changing the scale of the RTs by log-transforming them removes the interaction.* ***Middle****: a sub-additive interaction that, although not fully meeting Loftus' definition of "scale-independent" by crossing, is not necessarily destroyed by log-transformation.* ***Right****: a full cross-over interaction is least likely to be affected by assumptions about the scale. Study 4 focuses on the type of interactions described in the left and middle panels.*

Data and simulation conditions

All simulations in Study 4 assume the presence of one *large* main effect, but vary in whether there also is another *smaller* main effect and/or an *interaction* between the two main effects. For the simulations with non-zero interactions, we consider all three types of interactions shown in Figure 14: sub-additive, super-additive, and cross-over interaction (the scale-independent cross-over scenario corresponds to the presence of an interaction in the absence of the smaller main effect). Any non-zero effects generated in a simulation were either all generated to be linear in raw RTs, or all generated to be linear in log-RTs. For both effect scales, we compare the Type I error rates and power for the untransformed and log-transformed analysis approaches. The resulting simulation scenarios are summarized in Figure 15, and described in more detail next. To keep the number of simulation conditions feasible, we limit the simulations to the post-exclusion condition of the HS18 source data.

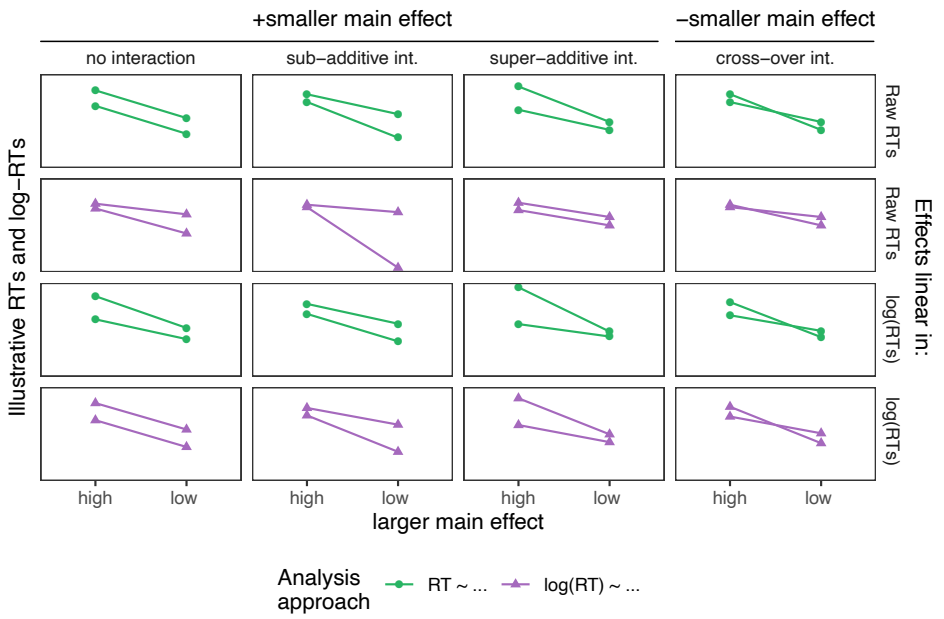

Figure 15. Schematic illustration of simulated smaller main effects and interactions in Study 4.

Generating non-null effects. We simulate the large main effect of almost 100 ms by splitting the HS18 data into early trials with large mean RTs (mean=395 ms, SD=204 ms) and late trials with small mean RTs (mean=299 ms, SD=136 ms). The two conditions of the large main effect are sampled from these two bins, so that half of each BATA's sampled items came from the bin with large mean RTs and half came from the bin with small mean RTs (for additional details, see Auxiliary Study 2d, SI-7).

The second (smaller) main effect and its interaction with the first main effect are generated following the same approach as in Study 2.[11] The smaller main effect is held identical to the medium effect (56 ms) used in Study 2. The interaction term is produced similarly, albeit with half the effect size of the smaller main effect (28 ms in raw RTs). The simulations in Study 4 thus describe the common scenario in which interactions are of smaller magnitude than the main effects, making them harder to detect.

[11] Alternatively, one could simulate both main effects by simply adding them. However, the implicit assumptions about the scale in which the effect is linear matter particularly much for larger effects (see Auxiliary Study 2d). We thus prefer an approach that takes the larger main effect from naturally occurring differences in RTs, guaranteeing that the data we bootstrap reflect the natural relation between RT means and RT variability.

Generating null effects. There are three scenarios of interest with null effects in Study 4: conditions in which the *interaction* term is null, conditions in which the smaller *main* effect is null (recall that the larger main effect is always present), and conditions in which both are null. Our Type I error simulations thus generate data under all three scenarios, and we analyze them separately below.

Effect scale. The smaller main effect and the interaction are always added in the same scale—i.e., either both linear in raw RTs or both linear in log-RTs. To add linear effects to raw RTs, we follow the same procedure as in Study 2. To add linear effects to log-RTs, we log the RTs, add the effect, and transform the resulting value back into raw RTs with the exponential function (for details, see SI-7.1). We note that this approach does *not* allow meaningful comparison of power across the two effect scales (raw or log-RTs). Meaningful comparison would require, at the very least, that the *marginal* effect size is held constant across the two effect scales, which is non-trivial (for further discussion, see footnote 23 in SI-7.1). Importantly, our approach does allow us to assess whether power results followed the same qualitative pattern as in Study 2 for each of the two scales.

Analysis approach. Following Study 2, we compare the untransformed and log-transformed analysis approach. For Study 4, the regression analyses included the full factorial terms of both main effects, as well as by-subject slopes for the large main effect. For Type I errors and power analyses, we focus on the smaller main effect and its interaction with the speed-up effect. We do not further focus on the larger main effect itself, which is always present in the Study 4 BATAs, and whose power is at ceiling under all simulation settings, regardless of analysis approach.

Results

Type I error rates. With two added effects (the smaller main effect and the interaction), Study 4 has two different effects to calculate Type I error rates for. The Type I results for the main effect are comparatively straightforward: no Type I error rates are significantly different from the expected value of 0.05 (Figure 16), although the untransformed analysis have significantly lower Type I error rates (for details, see SI-10.3).

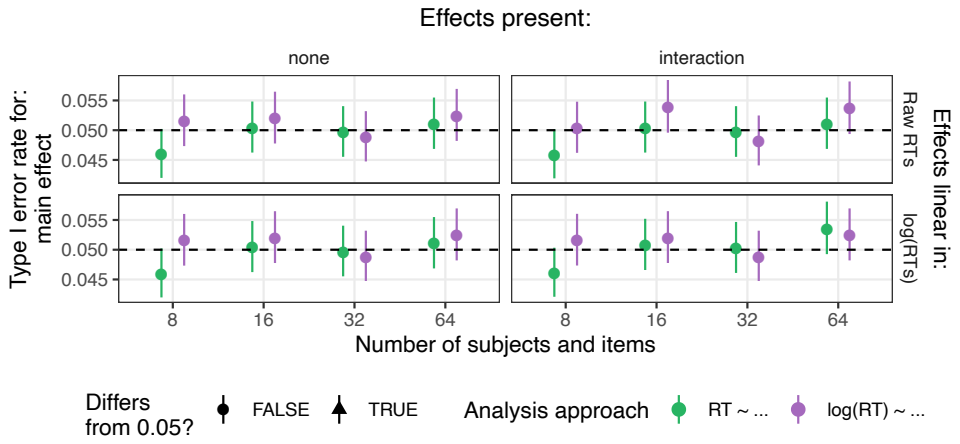

Figure 16. Type I error rates for the main effect in Study 4.

For the *interaction*, however, a more complex picture emerges (Figure 17). In the presence of a non-null (smaller) main effect, Type I error rates for the interaction are inflated. Specifically, the Type I error rates of analysis approaches that mismatch the scale in which the effect was generated increase with sample size: when the main effect is linear in raw RTs, the Type I error of the untransformed approach stays at the targeted .05, whereas the Type I error of the log-transformed approach increases up to almost twice the targeted rate for the largest data set; this qualitative pattern is flipped when the main effect is linear in log-RTs (for additional details, see SI-10.3).

To understand this pattern, consider that interactions assess whether the effect of one variable (e.g., the larger main effect) is independent of the effect of the other variable (e.g., the smaller main effect), i.e., whether the two effects are purely *additive*. If the smaller main effect is linear in raw RTs, the two main effects will be additive in raw RTs (i.e., no interaction). If analyzed as log-RTs, the two effects will, however, appear *sub*-additive (i.e., in log-RTs the difference between the two levels of the smaller effect will have a smaller difference when compared at the slower level of the larger effect than when compared at the faster level of the larger effect). In log-RTs, the two additive effects in raw RTs will thus appear as an interaction. And the larger the sample, the more likely this interaction effect will be to reach significance. Conversely, if the smaller main effect is generated to be linear in log-RTs, but then analyzed in raw RTs, the two effects will appear as a *super*-additive interaction (see "no interaction" panels in Figure 15). For

the log-transformed approach, the resulting anti-conservativity for interactions calls into question the across-the-board advantage of the log-transformed approach that we observed in the preceding simulations.

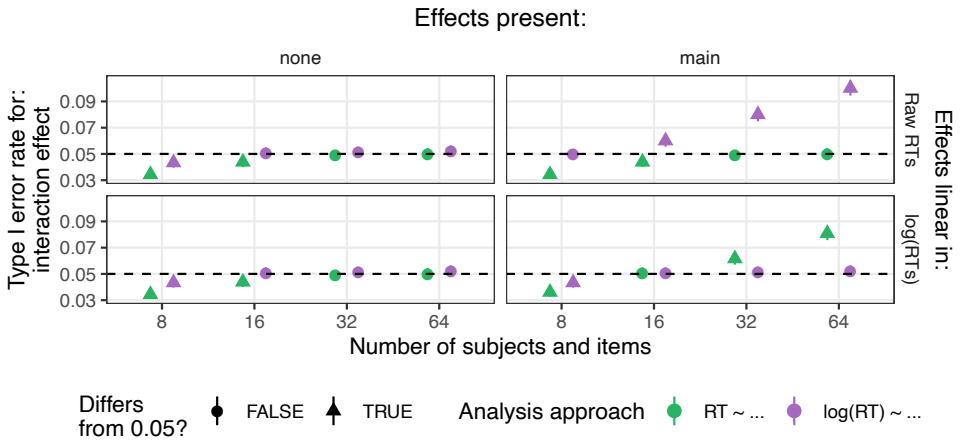

Figure 17. Type I error rates for the interaction effect in Study 4. When the smaller main effect was not present, the Type I error rates for the interaction effect were approximately 0.05, with the untransformed analysis approach having more conservative rates. However, when the smaller main effect was present, whether or not the main effect was linear in raw RTs or log-transformed RTs drastically affected which analysis approach had lower rates, with each approach performing better for the conditions that match its assumptions.

Power. Figure 18 visualizes the power (dis)advantage of the log-transformed analysis approach over the untransformed approach, for both the main effect and interaction (additional plots showing the power for each approach are provided in the SI-10.5).

For the main effect, the log-transformed analysis approach yields statistically indistinguishable or higher power than the untransformed approach in all simulation conditions (see SI-10.5), with an average power advantage of 0.26 in log-odds (a difference of 77.8% vs. 82% power). This replicates the pattern observed in the by-2 design of Study 2 (and Auxiliary Study 2d-e), which found a power advantage of the log-transformed approach even when effects were added to be linear in raw RTs. It is worth noting though that the presence of a super-additive interaction in log-RTs (blue line in the bottom-left panel of Figure 18) more or less nullifies this power advantage for the main effect.

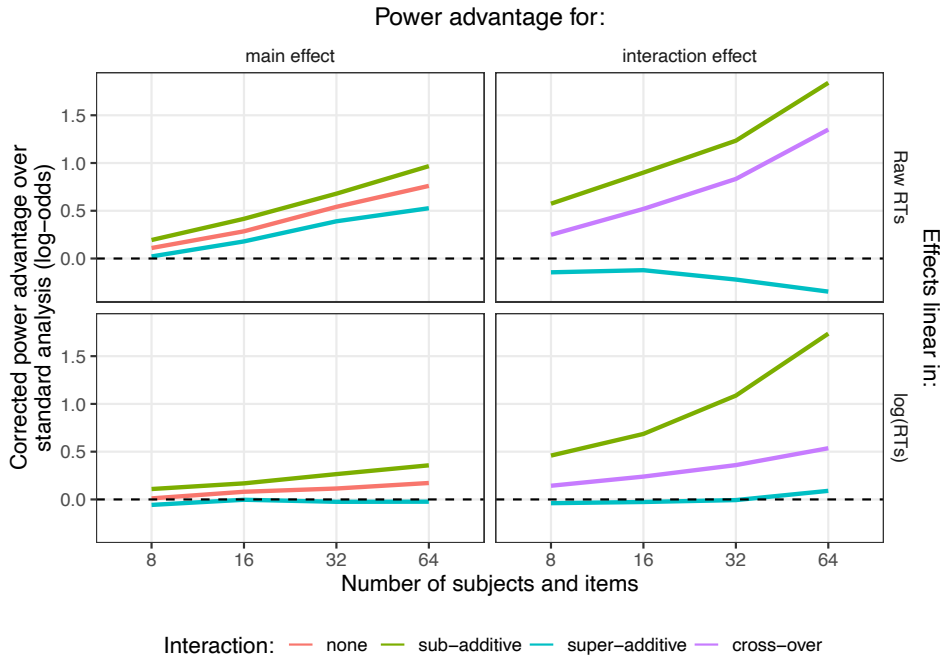

Figure 18. Relative power (dis)advantage of the log-transformed analysis in Study 4. Type I error-corrected relative power advantage of the log-transformed analysis approach compared to the untransformed analysis approach (as a difference in log-odds) for the interaction and main effect, depending on simulation condition. These results suggest that the primary deficit of the log-transformed analysis approach is when the full interaction is linear in raw RTs and is scale-dependent; in all other cases, the log-transformed approach does equivalently well or better than the untransformed analysis approach.

For the interaction, the pattern is again more complicated. For both cross-over and sub-additive interactions, the log-transformed approach always yields a power advantage, regardless of which scale the interaction was linear in. That sub-additive interactions are more often detected in the log-transformed approach makes intuitive sense: even when the interaction is actually sub-additive in raw RTs, analyzing this pattern in log-RTs will make it even *more* sub-additive, and thus easier to detect (see Figure 15). That the log-transformed approach also always seems to yield a power advantage for the cross-over interaction in Study 4 is more noteworthy. One possible explanation is that the simulations in Study 4 always contained a main effect that was much larger than the interaction. Future work is required to determine whether even *perfect* cross-over interactions (leading to the absence of main effects) yield a power advantage for the log-transformed approach.

Crucially, *super*-additive interactions did *not* always result in a power advantage for the log-transformed approach. As shown in the left panel of Figure 14, an interaction that is super-additive in raw

RTs can disappear when RTs are log-transformed, making it harder to detect. This explains the power *dis*advantage of the log-transformed approach for such scenarios. Like the power advantage for interactions that were generated in log-RTs, the power disadvantage of the log-transformed approach for interactions generated in raw RTs increases with sample size. While not as pronounced as the power advantage for the sub-additive interaction, the power disadvantage of log-transformed approach for the super-additive interaction is sufficiently large to matter for the range of designs we consider: for the largest BATAs, the power disadvantage approaches .9 log-odds, corresponding to a difference of, for example, 71.8% vs. 86.3% power—decreasing the researchers' odds to detect the interaction by a factor of almost 2.5.

Discussion

Some findings of Study 4 replicate Study 2, extending them to 2x2 factorial designs. We continue to find a power advantage of the log-transformed approach for main effects, and do so regardless of whether the true effects are linear in raw or log-RTs (see also Auxiliary Study 2d). We also find that this power advantage extends not only to sub-additive interactions, but also cross-over interactions, again regardless of which scale the effects are linear in—at least in the presence of one main effect that is larger than the cross-over interaction.

However, this general, *scale-independent* power advantage of the log-transformed approach does not extend to super-additive interactions: for such interactions, it matters whether the true (simulated) effects are linear in the scale for which the analysis assumes linearity. For super-additive interactions that are linear in raw RTs, the power advantage turns into a power *dis*advantage for the log-transformed analysis approach. This confirms and illustrates, for self-paced RTs, the *a priori* considerations for such interactions offered above and in previous work (Loftus, 1978; Wagenmakers et al., 2012). Of particular concern is that mismatches in the linearity assumption can also inflate the *Type I error* of untransformed and log-transformed analyses. For the designs simulated in Study 4, we find Type I error rates that are almost double the targeted Type I error rate when the analysis approach assumes linearity (of the smaller main effect) on the wrong scale (of the two scales considered in Study 4).

Additionally, all power advantages, all power disadvantages, and the Type I error inflation in Study 4 increased with sample size. This highlights that the specific magnitude of the observed (dis)advantages is of secondary relevance—it is expected to change with sample size, as well as the relative magnitude of the main effects and interactions. For example, general considerations about the logarithm-transform suggest that the Type I error inflation observed in Study 4 for interactions will further increase, depending on both the magnitude of the two main effects and overall mean reading speeds (for demonstration, see SI-11).[12] Of primary relevance here is that inflated Type I errors and power (dis)advantages can be observed for designs, sample sizes, and effect sizes that do occur in reading research.

What does this mean for researchers hoping to choose between analysis approaches? Researchers analyzing their experiment generally do *not* know which scale an effect is linear in *a priori* (unlike in the present simulation studies). The results of Study 4 thus mean that researchers cannot trivially determine whether a significant interaction in their experiment is more likely to be real or a Type I error. And researchers failing to find such interactions have no trivial way of determining whether the analysis suffered from a Type II error. Prior to additional considerations, it should matter little to those researchers that the log-transformed analysis has a power advantage in *most* scenarios simulated in Study 4, since it is unclear which simulation scenario their data come from! As we discuss next, this does not mean nothing can be concluded, and it certainly does not mean that the choice of analysis approach does not matter, or that the field can safely continue the status quo. Quite to the contrary.

[12] Similar reasoning suggests that the issues identified by Study 4 for log-transformed analyses would extend to log-shift analyses. In fact, compared to log-transformed analyses, log-shift analyses are expected to exhibit (1) even more inflated Type I error rates for interactions when effects are purely additive in raw RTs, and (2) an even stronger power disadvantage for interactions that are super-additive in raw RTs. These predictions follow from the fact that subtracting a constant from raw RTs (the shift of the log-shift model) before log-transforming RTs will *in*crease the difference of differences (i.e., the magnitude of the interaction) for those log-RTs, compared to not subtracting the constant.

General Discussion

Despite the central role that reading times continue to play across various fields, the ways in which they are typically analyzed make problematic assumptions. These assumptions—most notably, the assumption of normally distributed residuals with constant variance independent of the mean (homogeneity of variance)—have repeatedly been shown to be wrong for reaction times from simpler psychometric tasks (e.g., Baayen & Milin, 2010; Ratcliff, 1993; Rouder, 2005; Van Zandt, 2000; Wagenmakers & Brown, 2007). It thus seemed *a priori* plausible that the same assumptions would be violated by reaction times elicited from reading—a task that is arguably more complex than many of those previously studies, and generally assumed to tap into a broad range of mechanisms from ocular-motor control, to visual perception, to linguistic processing and cognitive control.

In Study 1, we found that this is indeed the case for the three source data sets we investigated. RTs in all three source data exhibited distributional properties that are unexpected under the assumption of normally distributed residuals. And, while log-transformed RTs to some degree matched those assumptions prior to censoring the data (outlier exclusion), this was no longer the case once censoring was applied. While similar issues—including the imperfect match of log-normally generated RTs—have previously been documented for reaction times from simpler tasks (e.g., Lo & Andrews, 2015; Rouder et al., 2005; Wagenmakers & Brown, 2007), most of these studies did not assess the *practical consequences* of these mismatching distributional assumptions (but see Liceralde & Gordon, 2022; Schramm & Rouder, 2019). This provided the motivation for the Type I error and power simulations of Studies 2-4.

In the remainder of this section, we first discuss the consequences of our findings for reading researchers. This leads us to discuss the importance of clearly stated assumptions in the interpretation of results, and motivates our recommendations for reading researchers:

> Especially when arguing based on the absence or presence of interactions (but also in some other cases we discuss below), reading researchers cannot in good conscience make scale-independent

conclusions based on a single analysis. Researchers should either clearly state why they assume effects to be linear in a particular RT-scale or conduct analysis over (at least) the most commonly assumed scales. If those analyses do not yield the same results, conclusions should clearly state this scale-dependence and discuss its possible interpretations.

Finally, we discuss the importance of assumptions for both power estimates, and comparisons of analysis approaches (as we did in the present study). With regard to the former, our recommendation is simple: the results of our simulation studies suggest that power estimates obtained under the assumption of normality or log-normality are so likely to be misleading that one might as well discard them.

Concrete takeaways for the simulated designs

We start by discussing the consequences of our studies *if they were to be taken at face value prior to questions about their generalizability*, which we address subsequently: what would be the take-home points for researchers who plan to analyze experiments with designs similar to those we simulated, and plan to do so—by choice or necessity—using LMM or similar analyses over power-transformed RTs (rather than more complex analyses)? We have in mind the majority of researchers in the cognitive, educational, and social sciences; researchers who are not statisticians, do not have access to super computers, but care about understanding their data and about reporting their results in ways that are reasonably likely to hold up to scrutiny.

Even when taking the assumption at face value, neither of the two analysis approaches we have considered in our studies provides a simple one-size-fits-all solution to reading time analyses. There are, however, some generalizations that emerge.

Designs with only a binary main effect. For designs with only a single main effect, the results of our studies seem to support a relatively simple conclusion (for the designs we simulated): the log-transformed approach has a clear and consistent advantage over the untransformed analysis, and—in line with arguments by Schramm and Rouder (2019)—analyses over log-shift transformed RTs performed even better. Our

findings suggest that researchers would face little downside when applying the log-shift approach to simple by-2 designs. Regardless of whether data censoring, residualization, or region aggregation was applied, the log-shift approach had a consistent power advantage in all three source data sets. Compared to the untransformed approach, this power advantage was often substantial (up to 41%)—in particular, for experiments that had more than negligible power to begin with—and did not come at the cost of inflated Type I errors.

Auxiliary Studies 2d and 2e replicated this general pattern, regardless of whether effects were linear in raw or log-RTs. The magnitude of the power advantage varied between simulation conditions, and did so systematically: for the designs considered in our studies, the power *advantage* of the log-transformed approach—when measured in log-odds—systematically increased with the power of the untransformed approach (see Figure 19B). Figure 19A suggests that the practical consequences of the choice between the two analysis approaches were most pronounced when the (actual) power of the untransformed approach fell between ~30-75%.

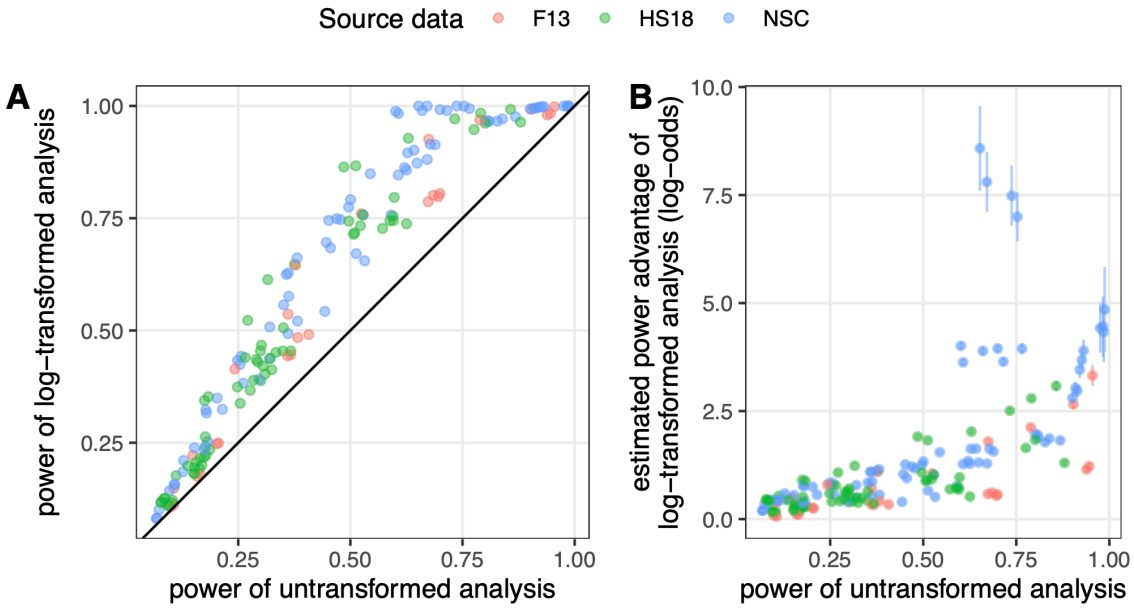

Figure 19. Comparing the Type I error-corrected power of the untransformed and log-transformed analysis approaches across all simulation conditions of Studies 2 and 2a-2e. ***Panel A:*** *power of both analysis approaches with a gray identity line.* ***Panel B:*** *power advantage (in log-odds) of the log-transformed analysis approach over the untransformed approach as a function of the untransformed analysis approach's power. Comparisons in which the log-transformed approach had 100% power are excluded. Intervals show*

95% CIs from logistic regression comparing the statistical power of two approaches. Future work is required to assess whether the pattern suggested by panels A and B generalizes beyond the source data and design types considered here. For example, the 'outliers' for the NSC data in Panel B seem to be due to particular high skew and kurtosis of the pre-exclusion NSC data (increasing with BATA size). This highlights how the advantages of an analysis approach can depend on the RT distributions of *the data submitted to analysis.*

Interaction designs. However, even for designs with comparatively simple 2x2 interactions (regardless of whether they are significant or not), Study 4 found that neither the untransformed, nor the log-transformed analysis approach affords scale-independent conclusions. When the actual scale of an effect mismatched that assumed by the analysis, both the untransformed and the log-transformed approaches yielded substantially inflated Type I error rates and wasted power, causing analyses to 'hallucinate' interactions or to miss genuine interactions (see also, Lo & Andrews, 2015; Loftus, 1978; Staub, 2021; Wagenmakers et al., 2012). These findings should concern researchers, given that interactions tests often play a critical role in testing theories.

While the issues we detected in Study 4 were limited to interactions and did not affect either of the two main effects, we note that this specific finding might be limited to the design decisions we made in Study 4. For example, compared to the main effects, the interaction effects were always small. On the one hand, this is a common scenario, since complete cross-over interactions are rare. On the other hand, it means that the simulations of Study 4 might have missed Type I error inflation and wasted power for main effects that would result when the interaction is larger than the main effect(s). We thus emphasize that the issue we identify lies with designs containing interactions, without necessarily being limited to the interactions itself.

Before we discuss the consequences of the scale-dependence, we note that this issue is *not* specific to reading studies, or reading times as dependent measure, but rather a general property of interactions. Neither is scale-dependence specific to comparisons of the identity- vs. log-transform: increased Type I error rates and reduced power for certain scale-dependent interactions are expected for *any* analysis approach that assumes linearity on a scale that is non-linearly (even if monotonically) related to the scale that the effects are actually linear in. For example, an inverse transform would also make additive effects in raw RTs appear sub-additive, while raw RTs would appear to have super-additive relationships if the effects were truly

additive in inverse RTs. In short, scale-dependence is an issue that exists independent of considerations about the nature of RT *distributions* (though distributional properties can ameliorate or increase the severity of scale-dependence).

Assumptions matter; they always have and always will. This means that researchers need to interpret their findings relative to, and conditional on, assumptions. This is, of course, true for *any* analysis. This means, among other things, that there is no safe "default" scale. And no method we are aware of completely frees researchers from consequences of scale assumptions (recall that even methods like generalized additive models, which allow non-linearity to emerge from the data, do so while penalizing against non-linearity on a scale assumed by the researchers). Our recommendations for reading researchers therefore are largely the same that apply to any analysis.

First, researchers should clearly state the assumptions they have made in their analysis, including the scale(s) they considered effects to be linear in. Whenever possible, these assumptions should be justified. In this context, we emphasize again that "ease of interpretation", which is sometimes evoked to justify a scale (e.g., Lo & Andrews, 2015; Osborne, 2002), is *not* a coherent argument: ease of interpretation matters little if the resulting interpretation is wrong. Theoretical motivations provide a better argument. For example, some theories commit to linking hypotheses that allow strong scientific inferences (in the sense of Platt, 1964). For instance, Smith and Levy's (2013) theory of word surprisal presents arguments as to why such effects are expected to be linear in raw RTs (but see Van Schijndel & Linzen, 2021). Of particular promise are cognitive or neural models with fully specified linking hypotheses that account for the generative process of RTs. Models such as SWIFT (Engbert et al., 2005), ACT-R (Lewis et al., 2013; Lewis & Vasishth, 2005), or EZ-READER (Reichle et al., 2003) are both theory-forward *and* capture aspects of RT distribution. These models let researchers commit to explicit, testable theories of RTs, and they allow the field to build on and revise those assumptions when necessary. Unfortunately, theories with clear linking hypotheses remain the exception. Similar to Loftus' appraisal of the memory literature at the time (1978), most studies that analyze RT or other psycholinguistic data continue to test informal, under-specified

hypotheses (for related arguments, Engelmann et al., 2019; Guest & Martin, 2021; Tanenhaus, 2004; Xie et al., 2023).

Second, to ameliorate the dependence on assumptions, we recommend that researchers validate their results for at least the most commonly employed scales. For RTs, we recommend that researchers replicate all analyses for at least raw and log-transformed (or better: log-*shift* transformed) RTs.[13] Open science platforms facilitate the publication of detailed supplementary information, and advances in freely available statistics software have made automatization of analysis easier than it ever was. Strong conclusions should only be drawn if they hold across different scales, or if clear arguments are given as to why one scale is to be preferred over the other (echoing Balota et al., 2013; e.g., see Kliegl et al., 2011; Staub, 2020, 2021). Sometimes, for example, an interaction—or a set of interactions across related RT measures—does not lend itself to a meaningful interpretation, and this can in and of itself form part of an argument as to why the effects should be interpreted on another scale (for demonstration, see Staub, 2020). We expect that adopting such practices will also deepen the researcher's *own* understanding of their data (e.g., what does it mean for an effect to be present in raw RTs but not log-RTs? for insightful discussion, see Staub, 2021).

We close this section with a clarifying note on a highly influential idea—that (1) additive effects on raw RTs indicate separate processing stages, whereas (2) interactions in raw RTs indicate that the interacting effects share a processing 'stage' (e.g., Roberts & Sternberg, 1993; Sternberg, 1969b, 2013).[14] This would seem to give increased importance to testing interactions in raw, rather than transformed, RTs. However, we do not see how Sternberg's considerations would free researchers from issues of scale-

[13] Alternatively or additionally, some researchers might prefer to include analyses over RTs that are transformed based on the results of a Box-Cox test applied to *their data* for *their* analysis formula (recall that the results of the Box-Cox test depend on the formula). We note, however, that (a) the Box-Cox test can recommend transformations that constitute highly implausible models of RTs (e.g., because they would suggest that RTs can be negative), that (b) arguably *none* of the power transformations considered by the Box-Cox test results in a *particularly* plausible model of RTs, so that one might doubt what value it adds to the analyses of RTs (see also Schramm & Rouder, 2019), and that (c) more normally distributed residuals—which is what the Box-Cox test assesses—do not *necessarily* entail better power and Type I error rates.

[14] Sternberg (1969b) himself often referred to 'stage' in single quotes, and highlighted the need to further define the notion. He also recognized the inherent bias of his proposal towards independent processing stages (since rejection of the hypothesis of serial, independent stages requires a significant interaction, whereas the lack of a significant interaction is taken as evidence against the hypothesis that two factors affect the same processing stage, p. 311).

dependence. Indeed, Sternberg himself clarified that additivity *supports,* but does *not imply*, separate processing stages (Sternberg, 1969b, 2013). He was careful to point out that "one can imagine exceptions [to both (1) and (2)]" (Sternberg, 1969b, p. 282). For example, indirect effects can create interactions across processing stages (ibid), or parallel processes at the same processing stage can lead to interactions without a shared processing 'stage' (p. 309-310). Two factors might even happen to have additive effects on the *same* processing stage (e.g., under certain assumptions, ideal observer models predict additive effects of input and context, Bicknell et al., 2016, p. 282). Perhaps most importantly, Sternberg *took for granted* that RTs should not be transformed, since RTs are "a basic measure" (Sternberg, 1969b, p. 311; see also pp. 286-7). However, not all effects are linear in raw RTs; and even when effects are linear over *some* range of raw RTs, this linearity might break down when a larger range of RTs is considered. A careful application of Sternberg's idea thus requires researchers to first assess that the effects in question are linear in raw RTs over the entire range of RTs resulting from the cumulative effects of the potentially interacting factors (which Sternberg himself did for several phenomena he studied, e.g., Sternberg, 1969a).

Scale-dependence is *not* specific to linear models

As mentioned in Study 4, issues of scale-dependence are not specific to linear models. We anticipate, for example, that most of the advanced analysis methods discussed in the introduction will not preempt scale-dependence (though empirical advances that shed light on which scale effects on reading tend to be linear in might). Previous work has compared large range of theoretical distributions against properties of natural RT distributions, such as the correlation between RT means and standard deviations (e.g., the Wald, Gamm, Weibull, log-normal, etc., Palmer et al., 2011; Van Zandt, 2002; Wagenmakers & Brown, 2007). This includes analyses approaches that also model variance, skew, or other parameters of the outcome distribution, rather than just the mean (or location), such as distributional regression or similar approaches (among others: Balota & Yap, 2011; Van Zandt, 2000, 2002). Such approaches can account for, e.g., correlations between the mean and variance of RTs, which would otherwise violate the assumption of homogeneity. Other analysis approaches have focused on describing RTs as the result of multiple processes,

typically as a mixture of two components (e.g., log-shift, Nicenboim et al., 2014; Nicenboim & Vasishth, 2018; ex-Wald, Schwarz, 2001; ex-Gaussian, Staub & Benatar, 2013; Vasishth et al., 2018). Of these, the log-shift model—also in light of Studies 1 and 2—might hold particular promise, as its parameters link more transparently to models of perceptual decision-making (for discussion, see Kieffaber et al., 2006; Moutsopoulou & Waszak, 2012; Wagenmakers & Brown, 2007).

However, while log-shift and similar models might better capture the distributional properties of RTs, they do not preempt scale-dependence of interactions. For example, while log-shift models better capture the positive soft lower bound of RTs (and might even ameliorate some of the severity of the issues found in Study 4), additive effects in raw RTs would still be expected to appear as interactions at a higher-than-targeted Type I error rate in log-shifted RTs.

Generalized additive mixed-effect models (GAMMs, Wood, 2011) constitute a potentially more promising tool to deal with scale assumptions. As we discussed in Study 4, even GAMMs are affected by the scale of the outcome variable, since they penalize against non-linearity on the scale assumed by the researcher. However, GAMMs at least allow researchers to assess whether interactions still hold when the main effects are allowed to exhibit non-linear effects on the scale assumed by the researchers. We see two important limitations of this approach. First, it does not lend itself to categorical predictors, and thus cannot easily be applied to the designs we have discussed here. Second, while GAMMs can be a powerful diagnostic or exploratory tool, they come with many additional researcher degrees of freedom. This can increase the risk of overfitting (especially when the experiment was not explicitly designed with a GAMM analysis in mind). We thus would not recommend this approach as the primary analysis approach for non-experts.[15]

[15] An alternative, which we have not seen applied to RT analyses, is rank regression (Cuzick, 1988; Iman & Conover, 1979, not to be confused with rank-based estimation of regression). Rank regression predicts the mean rank of outcomes, rather than mean of the outcome, and was developed specifically with the purpose of reducing scale dependence. It can be applied to factorial designs and provides fewer researcher degrees of freedom. It is, however, not without downsides: by reducing outcome observations to their rank, rank regression discards information about the outcome, which reduces its sensitivity and statistical power.

How likely are our results to generalize to *your* data?

The results we have summarized above are based on three source data sets that reflect different approaches to the elicitation of self-paced RTs (lab vs. web), different types of sentence stimuli (isolated vs. connected discourse), etc. Together, these source data sets contained more than 1 million unique word RTs from over 600 sentence items read by more than 500 subjects. In total, the studies we conducted are based on over 4.65 million BATAs, each resampling data sets with 64 to 4,096 observations (depending on the simulated sample size). The power advantage for main effects replicated when we sampled subsets of the data with different mean RTs, different standard deviations, different degrees of skew and kurtosis; they replicated for between- and within-subject and -item designs; and they replicated when we generated data with additional random by-subject and -item variability (regardless of whether that variability was linear in raw or log-RTs).

At the same time, all of our simulations are limited to (subsets of) *only* three source data sets. While the inclusion of web-based source data means that the bootstrapped RTs reflect reading from a broader population than college students, all of these source data came from self-paced reading, rather than eye-tracking reading, paradigms; all of the subjects were recruited to be self-reported native speakers of the same WEIRD language (English); none of the subjects were recruited from populations with known reading deficits; the sentence items in the source data are unlikely to reflect the full range of linguistic diversity found in English texts (or even psycholinguistic studies); etc. One should therefore not blindly generalize our findings to other data sets. It is unclear, for example, whether RTs from eye-tracking reading studies exhibit the same distribution as self-paced RTs (though we are not aware of findings that suggest otherwise for, e.g., first-pass or total reading times; see also side-by-side comparisons of RTs from the two reading paradigms in e.g., Shain & Schuler, 2021; Smith & Levy, 2013).

Our simulation conditions, too, were limited in scope: the most complex design we simulated was a 2x2 design with one large main effect, a smaller (or null) main effect, and an even smaller (or null) interaction. On the one hand, we consider it plausible that some of our findings generalize beyond the specific designs we simulated. We see no obvious reason, for example, why the power advantage for main

effects in simple by-*N* designs should fail to extend to the detection of main effects with $N > 2$ levels. However, even relatively simple extensions of such designs or analyses might require additional considerations. For example, the *comparison* of effect *sizes* at different RT means—such as the comparison of successive differences in $N > 2$ condition means—is expected to suffer from the same scale-dependence that we identified for interactions. Similarly, it is less clear to us whether the scale-independent power advantage of the log-transformed or log-shift approaches that we observed for by-2 designs extends even to simple analyses with a single *continuous* predictor (as is the case, e.g., in many studies on surprisal, Linzen & Jaeger, 2014; Smith & Levy, 2013; Van Schijndel & Linzen, 2021). And even though we expect the general issue we identified for interactions to extend to (and possibly be exacerbated in) higher-order interactions, we suspect that at least the severity of the issue will depend on the size of the interaction relative to the size of the interacting effects.

Examples like these highlight the need to be conservative in generalizing the findings of the present studies beyond the designs we simulated. They also highlight the need for further simulation studies. The hierarchical bootstrap approach can be extended to many of these questions described above, providing a way for future work to assess whether our findings extend to other source data, other designs, or even effects on other scales.

Effects of assumptions on power (and other) estimates

The importance of considering assumptions is not limited to the *analysis* of reading data. It equally applies to estimates of Type I error, statistical power, precision, and other properties of analyses. In Study 3, we found that Type I error and power estimates based on normally or log-normally generated RTs can differ drastically from estimates based on bootstrapped natural RTs. Most concerningly, the results of Study 3 suggest that the most commonly employed approaches to power estimation can yield power estimates that are so inflated for RT analyses that they are meaningless. As we spell out next, this has far-reaching consequences for the field, both for future work and for the interpretation of past work.

Consider the scenario that applies to the vast majority of power estimates reported in the reading literature: researchers report power that is estimated under the exact same assumptions that hold for the analysis they plan to apply (or already have applied) to their data. This is, for example, what default application of the popular R package *simr* (Green & MacLeod, 2016) will do: researchers fit an LMM to their data—e.g., over raw or log-RTs—and then ask *simr* to calculate the statistical power of their analysis.[16] For such power analyses, *simr* and similar software will repeatedly generate data under the assumption of the LMM—i.e., under the assumption of normally distributed raw or log-RTs—and then analyze the generated data with the exact same LMM that the data was generated from (see also Brysbaert & Stevens, 2018; Van Zandt, 2002).

In Study 3, we found that this approach can *massively inflate* power estimates for RT analyses, leaving researchers with a false sense of security. For example, for an experiment with 32 subjects, 32 critical items, and a "medium" effect, simulations that generated normally distributed RTs and then analyzed the generated RTs under that same assumption of normality (Figure 12a) estimated statistical power above 80%. Bootstrap simulations over natural RTs instead found power to be below 35%! Similarly inflated power was observed when RTs were both generated and analyzed under the assumption of log-normality. Indeed, power estimates in Study 3 were inflated, even when the parametric assumption under which RTs were generated *mis*matched the parametric assumptions of the analysis. For example, even when RTs were generated under the assumption of log-normality, power estimates were substantially inflated compared to those obtained for bootstrapped RTs (see Figure 12c). One possible reason for this is suggested by Study 1: even log-normally generated RTs deviate less from the assumption of normality than actual RT distributions.

[16] For such power simulations, *simr* will alert users that they are conducting a post-hoc power analysis (which suffers from issues beyond those discussed here). Alternatively, users can specify one model for data generation and another model for data analysis. But even if data generation and analysis make different analytic assumptions, it is not guaranteed that this captures the issues that would arise for natural RTs (see Figure 12c in Study 3). We emphasize that these issues pertain to how *simr* and similar software is *applied and interpreted by researchers*, not with the software itself.

Whatever the reasons for the inflated power, we submit that parametric power estimates assuming normality or log-normality of RTs are close to meaningless, and should be avoided in future work. Where arguments based on power estimates have been critical in the evaluation of past work, those arguments should be revisited. Most obviously, this applies to seemingly high-powered replication failures (though those replications will tend to still have *more* power than the original study, even if their actual power might be nowhere close to that reported). Under-powered experiments do, however, also risk inflated Type I error rates (Simmons et al., 2016).

Moving forward, one option available to researchers is to obtain power estimates under a broad range of parametric assumptions (going beyond normality and log-normality of RTs). Another option is to obtain power estimates through bootstrap. While assessments of analysis approaches—as in the present study—are computationally costly because they have to span a large range of possible scenarios, bootstrap power simulations for a single planned analysis can be conducted on modern laptops. The hierarchical bootstrap approach employed in our studies, and the R code we share as part of our OSF repository, provide one path towards such analyses. Researchers interested in this approach will face a particular challenge: as already mentioned, bootstrap and similar non-parametric approaches to power simulations make the strong assumption that the source data is representative of the data one seeks to analyze. Just as the assumption of parametric data generation approaches can make power estimates uninformative, so too can the choice of non-representative source data for the bootstrap.

For post-hoc power analyses, the assumption of the representativeness largely reduces to the assumption that the source data is sufficiently large, since researchers can bootstrap the data they collected in the experiment. However, this approach is not possible for power analyses that are conducted to guide experimental design, participant recruitment, etc. *prior* to the experiment—i.e., the type of power analysis that should be the gold standard. In addition to finding sufficiently large available source databases, researchers who seek to use bootstrap for such power simulations will have to be particularly careful in selecting source data that can be reasonably expected to be representative of the data they aim to collect in their experiments. Any effects of design, procedure, participants, stimulus materials, etc. on the RTs in the

source data will be inherited by the bootstrapped data. Once the data is collected, the assumption of representativeness can be assessed by comparing the distributional properties of the collected data and the source data power was estimated from.

Effects of assumptions on analysis recommendations

Finally, assumptions also play a critical role in understanding the results of previous work on the *reliability of analysis approaches* for non-reading reaction time data (e.g., Liceralde & Gordon, 2022; Schramm & Rouder, 2019), as we have done here for reading data. Both of these previous studies employed parametric data generation to estimate Type I error rates and power of different analysis approaches. For such studies, the over-estimation of absolute statistical power described in the previous section is not necessarily an issue. Instead, these studies aim to compare the *relative* power (and other properties) of different analysis approaches. Critically, the results of Study 3 suggest that this question, too, can be unduly affected by parametric assumptions about the distribution of RTs. We found that the power (dis)advantage of the log-transformed analysis approach depended on the assumptions under which RTs were generated. This raises questions about whether similar issues apply to previous work on the reliability of reaction time analyses.

In one of the most comprehensive simulation studies on the reliability of reaction time analyses, Liceralde & Gordon (2022) used LMMs fit to raw or power-transformed reaction times to generate new data. Going beyond our Study 3, Liceralde and Gordon further aimed to capture the distributional properties of reaction times by assuming trial-level residuals to be Gamma distributed. Schramm & Rouder (2019) instead use a generative model of perceptual decision-making—the shifted, one-bound diffusion model—to generate the reaction times. Based on the results of their Type I error and power simulations, both studies recommend the use of LMMs over raw, rather than power-transformed, reaction times. However, unlike

the present study, neither study compared the distributional properties of the generated reaction times against the distribution of natural reaction times.[17]

Without such comparisons, it is unclear whether the distribution of generated reaction times in those studies reflected the distribution of natural reaction times. For future research that seeks to make recommendations for reaction or reading time analyses, we thus suggest that researchers complement parametric simulations with non-parametric approaches to data generation like the hierarchical bootstrap. Similar recommendations arguably apply for any other outcome measures that are not known to follow the assumptions of parametric approaches to data generation. In the same vein, we recommend that analysis recommendations are based on comparison of Type I error rates and power (or their Bayesian extensions), rather than 'only' comparisons of distributional properties (e.g., Anders et al., 2016; Palmer et al., 2011; Schwarz, 2001; Wagenmakers & Brown, 2007), since even subtle deviations from the natural distribution of RTs might affect the reliability of analysis approaches.

Conclusion

Using an under-explored non-parametric hierarchical bootstrap approach to generate naturally distributed reading time data, we find the two most common approaches to reading time analyses make assumptions that mismatch the actual distributional properties of RTs. Critically, we also find that these mismatching assumptions have *practical* consequences for the statistical power of and, in some cases, the Type I error of analyses. Unfortunately, these negative consequences can be easily missed by the type of parametric approach to power or Type I error simulation that remain most common in the field. For researchers that seek to assess the statistical power of their own analyses, this means that parametric simulation approaches can easily provide a false sense of security, sometimes *substantially* over-estimating

[17] Additionally, Liceralde & Gordon (2022) did not include simulations with log-normally generated data—i.e., the power-transform-based data generation approach that we find to most closely approximate RT distributions prior to exclusions in Study 1. Liceralde and Gordon do not consider this approach because Box-Cox tests on their source *reaction* time data did not suggest the log-transform ($\lambda = 0$).

statistical power. We hope that the R code provided as part of the OSF repository for our studies provides a helpful starting point to those who seek to estimate the Type I error, power, or other properties of their analyses.

Acknowledgements

We are grateful to Wednesday Bushong, Marc Brysbaert, Shravan Vasishth, Adrian Staub, and an anonymous reviewer for constructive and insightful feedback on earlier versions of this article. We also owe many thanks to Van Liceralde for helpful conversations about both our and their project (Liceralde & Gordon, 2022). This work grew out of some initial simulations reported in an appendix of Jaeger, Bushong, & Burchill, 2019. Earlier versions of this work were presented at the 2018 AMLaP conference.

References

Anders, R., Alario, F., & Van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. *Psychological Methods*, *21*(3), 309.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036.

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

Baayen, R. H., Vasishth, S., Bates, D. M., & Kliegl, R. (2016). The Cave of Shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. https://doi.org/10.1093/jnci/85.5.365

Balota, D. A., Aschenbrenner, A. J., & Yap, M. J. (2013). Additive effects of word frequency and stimulus quality: the influence of trial history and data transformations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1563.

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. *Current Directions in Psychological Science*, *20*(3), 160–166.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiv Preprint ArXiv:1506.04967*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bengtsson, H. (2019). *future: Unified Parallel and Distributed Processing in R for Everyone*.

Bicknell, K., Jaeger, T. F., & Tanenhaus, M. K. (2016). Now or ... later: Perceptual data is not immediately forgotten during language processing. *Behavioral and Brain Sciences*, *39*, 23–24.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1).

Carpenter, R. H. S., & Williams, M. L. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature*, *377*(6544), 59–62.

Cuzick, J. (1988). Rank regression. *The Annals of Statistics*, 1369–1389.

D'Agostino, R. B. (1970). Transformation to normality of the null distribution of g1. *Biometrika*, 679–681.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Ehrlich, K., & Rayner, K. (1983). Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, *22*(1), 75–87.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: a dynamical model of saccade

generation during reading. *Psychological Review*, *112*(4), 777.

Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, *43*(12), e12800.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS One*, *8*(10).

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2018). The Natural Stories Corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Green, P., & MacLeod, C. J. (2016). simr: an R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: an example using the Stroop task. *Psychological Bulletin*, *109*(2), 340.

Heider, P. M., Dery, J. E., & Roland, D. (2014). The processing of it object relative clauses: Evidence against a fine-grained frequency account. *Journal of Memory and Language*, *75*, 58–76.

Hinkley, D. (1994). [Bootstrap: More than a Stab in the Dark?]: Comment. *Statistical Science*, *9*(3), 400–403.

Iman, R. L., & Conover, W. J. (1979). The use of the rank transform in regression. *Technometrics*, *21*(4), 499–509.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jaeger, T. F., Bushong, W. R., & Burchill, Z. (2019). *Strong evidence for expectation adaptation during language understanding, not a replication failure. A reply to Harrington Stack, James, and Watson (2018)*.

Jegerski, J. (2014). Self-paced reading. In J. Jegerski & B. VanPatten (Eds.), *Research methods in second language psycholinguistics* (pp. 20–49).

Kieffaber, P. D., Kappenman, E. S., Bodkins, M., Shekhar, A., O'Donnell, B. F., & Hetrick, W. P. (2006). Switch and maintenance of task set in schizophrenia. *Schizophrenia Research*, *84*(2–3), 345–358.

Kliegl, R., Masson, M. E. J., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition*, *18*(5), 655–681.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, *1*, 238.

Knief, U., & Forstmeier, W. (2018). Violating the normality assumption may be the lesser of two evils. *BioRxiv*, 498931.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). {lmerTest} Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26.

https://doi.org/10.18637/jss.v082.i13

Lachaud, C. M., & Renaud, O. (2011). A tutorial for analyzing human reaction times: How to filter data, manage missing values, and choose a statistical model. *Applied Psycholinguistics*, *32*(2), 389–416.

Lewis, R. L., Shvartsman, M., & Singh, S. (2013). The adaptive nature of eye movements in linguistic tasks: How payoff and architecture shape speed-accuracy trade-offs. *Topics in Cognitive Science*, *5*(3), 581–610.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Liceralde, V. R. T., & Gordon, P. C. (2022). Consequences of Using Power Transforms as a Statistical Solution in Linear Mixed-Effects Models of Chronometric Data. *Vanderbilt University*.

Linzen, T., & Jaeger, T. F. (2014). Investigating the role of entropy in sentence processing. *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, 10–18.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171.

Loftus, G. R. (1978). On interpretation of interactions. *Memory \& Cognition*, *6*(3), 312–319.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization* (Issue 8). Oxford University Press on Demand.

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*(5), 861–904.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin \& Review*, *16*(5), 798–817.

Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. *New Methods in Reading Comprehension Research*, 69–89.

Moutsopoulou, K., & Waszak, F. (2012). Across-task priming revisited: response and task conflicts disentangled using ex-Gaussian distribution analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(2), 367.

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34.

Nicenboim, B., Vasishth, S., & Kliegl, R. (2014). Readers with less cognitive control are more affected by surprising content: Evidence from a self-paced reading experiment in German. *IEICE Technical Report; IEICE Tech. Rep.*, *114*(176), 67–71.

Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics*, *40*, 26–55.

Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation*, *8*(1), 6.

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58.

Platt, J. R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347–353.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*.

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

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333.

Ratcliff, R., & Van Dongen, H. P. A. (2011). Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proceedings of the National Academy of Sciences*, *108*(27), 11285–11290.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–476.

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, *14*, 611–653.

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*(2), 377.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223. https://doi.org/10.3758/BF03257252

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *Journal of the Royal Statistical Society Series D: The Statistician*, *41*(2), 169–178.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*(5), 309–316.

Schramm, P., & Rouder, J. (2019). Are Reaction Time Transformations Really Beneficial? *PsyArXiv*.

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, \& Computers*, *33*(4), 457–469.

Shain, C., & Schuler, W. (2021). Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition*, *215*, 104735.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2016). *False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.*

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments, \& Computers*, *28*(4), 516–536.

Stack, C. M. H., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877.

Staub, A. (2020). Do effects of visual contrast and font difficulty on readers' eye movements interact with effects of word frequency or predictability? *Journal of Experimental Psychology: Human Perception and Performance*, *46*(11), 1235.

Staub, A. (2021). How reliable are individual differences in eye movements in reading? *Journal of Memory and Language*, *116*, 104190.

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

Stephen, D. G., & Mirman, D. (2010). Interactions dominate the dynamics of visual cognition. *Cognition*,

115(1), 154–165.

Sternberg, S. (1969a). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457.

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

Sternberg, S. (2013). The meaning of additive reaction-time effects: Some misconceptions. *Frontiers in Psychology*, *4*, 744.

Stone, K., von der Malsburg, T., & Vasishth, S. (2020). The effect of decay and lexical uncertainty on processing long-distance dependencies in reading. *PeerJ*, *8*, e10438.

Tanenhaus, M. K. (2004). On-line sentence processing: past, present and, future. *On-Line Sentence Processing: ERPS, Eye Movements and Beyond*, 371–392.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166.

Van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin \& Review*, *7*(3), 424–465.

Van Zandt, T. (2002). Analysis of response time distributions. *Stevens' Handbook of Experimental Psychology*, *4*, 461–516.

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PloS One*, *8*(10), e77006.

Vasishth, S., Mertzen, D., JÃ€ger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. https://doi.org/https://doi.org/10.1016/j.jml.2018.07.004

Vaughan, D., & Dancho, M. (2018). *furrr: Apply Mapping Functions in Parallel using Futures*.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer.

Wagenmakers, E. J., & Brown, S. (2007). On the Linear Relation Between the Mean and the Standard Deviation of a Response Time Distribution. *Psychological Review*, *114*(3), 830–841. https://doi.org/10.1037/0033-295X.114.3.830

Wagenmakers, E. J., Farrell, S., & Ratcliff, R. (2005). Human cognition and a pile of sand: a discussion on serial correlations and self-organized criticality. *Journal of Experimental Psychology: General*, *134*(1), 108.

Wagenmakers, E. J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory \& Cognition*, *40*(2), 145–160.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, *72*, 37–48.

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex*.

Yan, S., & Jaeger, T. F. (2020). Expectation adaptation during natural reading. *Language, Cognition and Neuroscience*, *35*(10), 1394–1422.

Supplementary Information

for Burchill & Jaeger

How reliable are standard reading time analyses?

Hierarchical bootstrap reveals substantial power over-optimism and scale-dependent Type I error inflation

SI-1 The soft lower bound of reading times

In the main text, we refer to the soft lower bound of RTs. This bound is strikingly visible, for example, in the changes in RTs across an experiment shown in Figure 20: as participants become more familiar with a self-paced reading (SPR) task, their RTs often move closer to their lower bound.

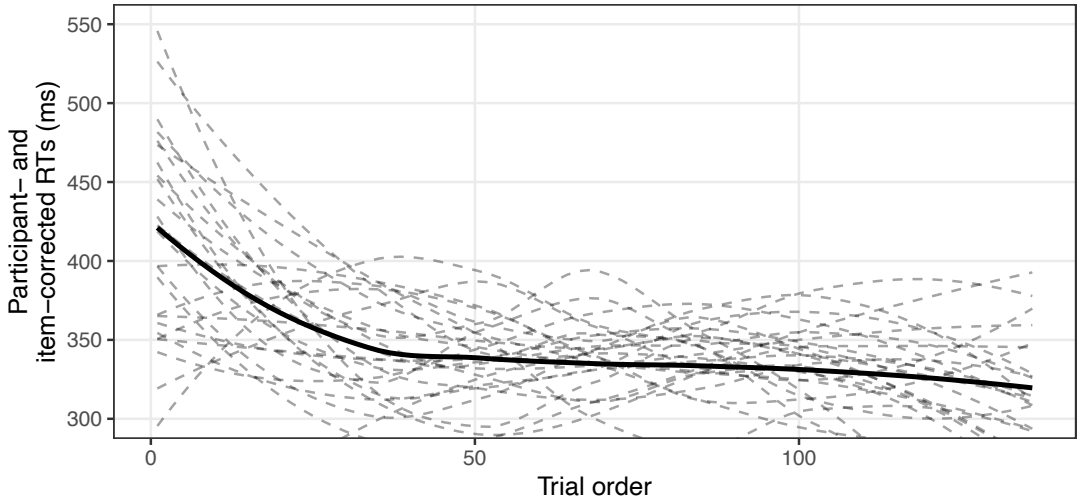

Figure 20. Changes in RTs across trials (here shown for NSC data from Study 1). As subjects become familiar with the SPR task over the course of an experiment, their RTs approach a lower bound. To capture the (non-linear) effect of trial on RTs, we used a generalized additive mixed model with non-linear predictors for within-sentence word order and trial order, with random intercepts by subjects and sentence items. The solid black line is a smooth of the predicted RTs from this model, marginalizing over subjects and items. The dashed lines are smooths of 30 randomly selected subjects' RTs, after residualizing out the random by-subject and by-item effects from the model.

SI-2 Study 1

This section provides additional result plots for the F13 and NSC data that parallel those provided in the main text for HS18.

SI-2.1 Distributional properties of source RTs

Figure 21 and Figure 22 replicate Figure 1 from the main text for the F13 and NSC data, respectively. Figure 23 and Figure 24 replicate Figure 2 for the F13 and NSC data, respectively.

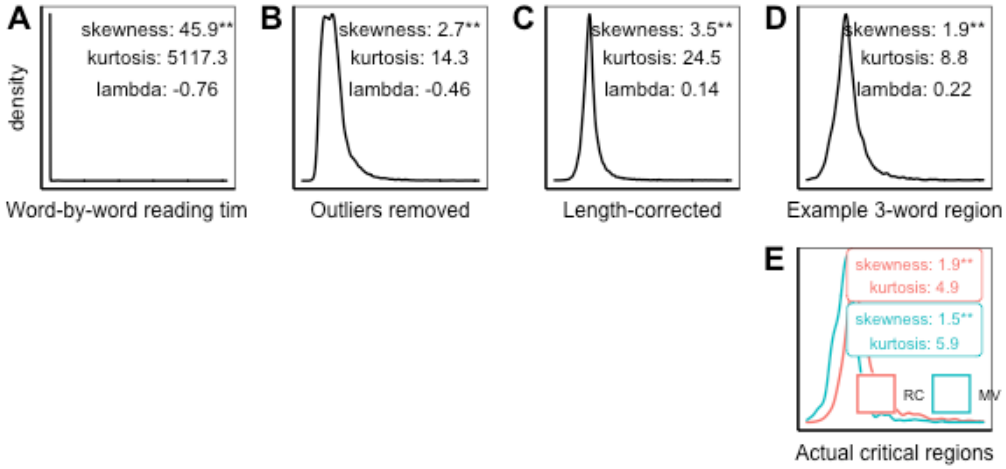

Figure 21: Marginal density of word-by-word reading times (RTs) using filler-trial data from Fine et al. (2013), Experiment 2. ***Panel A:*** *Raw RTs.* ***Panel B:*** *Raw RTs after outlier exclusion of RTs ≤ 100 ms or ≥ 2000 ms.* ***Panel C:*** *Length-corrected RTs after outlier exclusions, which also correct for individual differences in reading speeds between participants.* ***Panel D:*** *Average length-corrected RTs for a randomly chosen three-word region.* ***Panel E:*** *The length-corrected RTs averaged across the critical regions analyzed in the original study (Harrington Stack et al., 2018). Skewness is calculated as* $E[((X-\mu)/\sigma)^3]$ *and the kurtosis is the excess kurtosis calculated via Pearson's measure of kurtosis,* $E[((X-\mu)/\sigma)^4]$*. All distributions are significantly skewed based on D'Agostino test (D'Agostino, 1970). Box-Cox λs indicate which power transform would make the distribution of residuals most normal (see text for detail).*

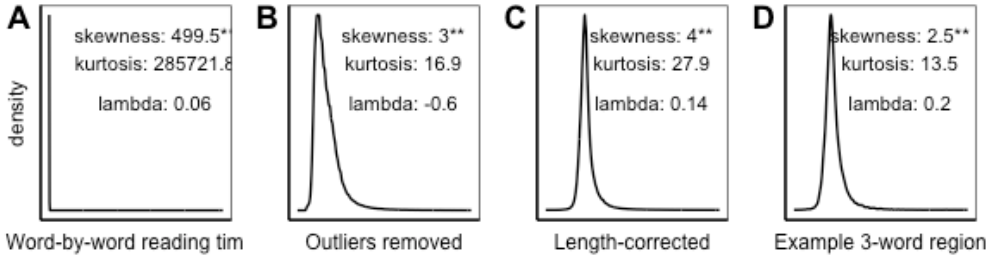

Figure 22: Marginal density of word-by-word reading times (RTs) using filler-trial data from the Natural Stories Corpus (Futrell et al., 2017), Experiment 2. ***Panel A:*** *Raw RTs.* ***Panel B:*** *Raw RTs after outlier exclusion of RTs ≤ 100 ms or ≥ 2000 ms.* ***Panel C:*** *Length-corrected RTs after outlier exclusions, which also correct for individual differences in reading speeds between participants.* ***Panel D:*** *Average length-corrected RTs for a randomly chosen three-word region. Unlike the HS18 and F13 datasets, the NSC data is from a reading corpus without a factorial manipulation (hence no Panel E).*

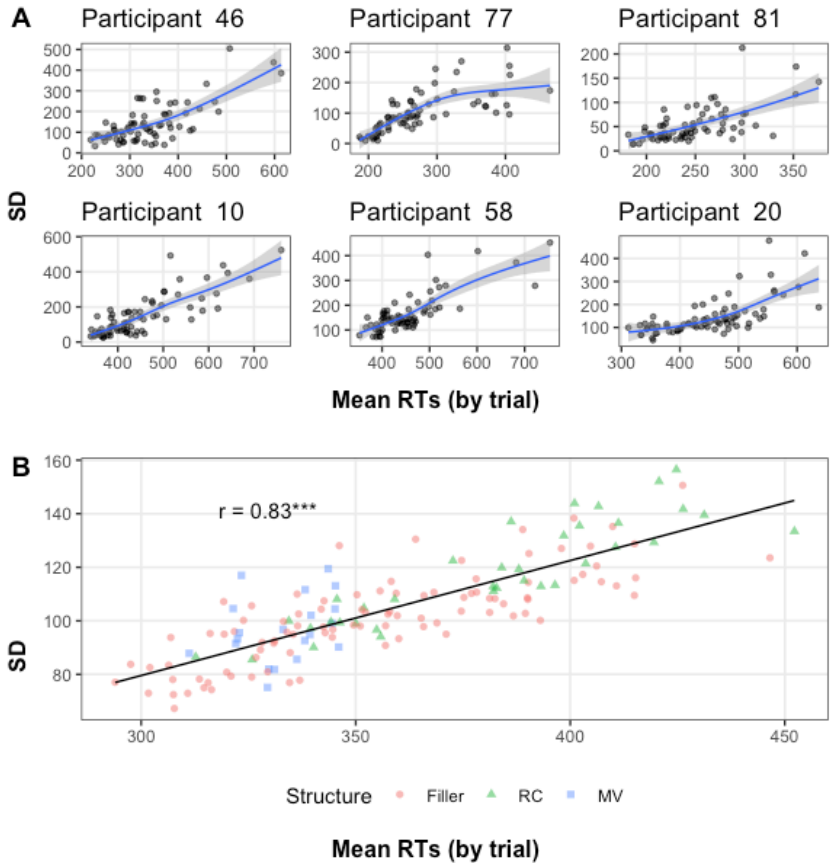

Figure 23: Correlation between means and standard deviations (SD) of outlier-excluded word-by-word reading times in the F13 data. ***Panel A:*** *Randomly drawn participants from the F13 data. Points reflect mean sentence RTs. Note that axis limits vary across panels.* ***Panel B:*** *Correlation across sentence items. Each data point represents one sentence. Shape and color show the sentence structure. RT means are a highly significant linear predictor of SDs ($p<0.001$).*

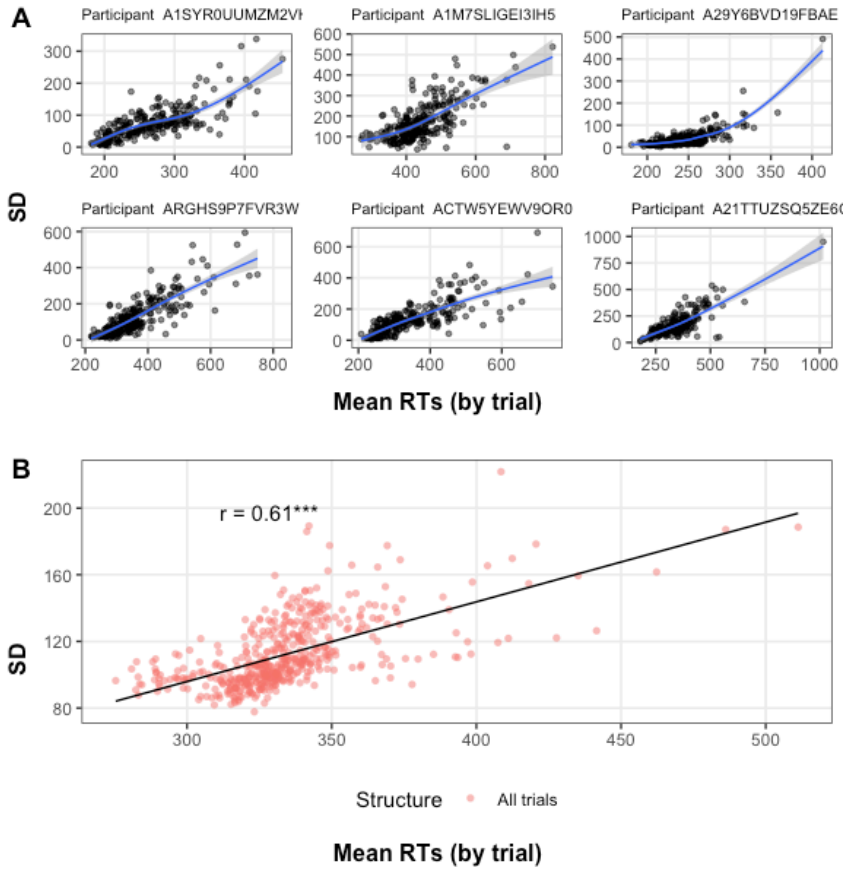

Figure 24: Correlation between means and standard deviations (SD) of outlier-excluded word-by-word reading times in the NSC data. ***Panel A:*** *Randomly drawn participants from the NSC data. Points reflect mean sentence RTs. Note that axis limits vary across panels.* ***Panel B:*** *Correlation across sentence items. Each data point represents one sentence. Shape and color show the sentence structure. RT means are a highly significant linear predictor of SDs ($p<0.001$).*

SI-2.2 Additional information on the hierarchical bootstrap procedure

The hierarchical bootstrap procedure differed somewhat between, on the one hand, the two source data that originated from factorial experiments (HS18 and F13) and, on the other hand, the NSC corpus. For both types of source data, we describe how we sampled BATAs depending on the four data preparation conditions. We first describe the approach for the HS18 and F13 data, and then describe how the procedure for the NSC data differed.

Pre-exclusion BATAs. First, we extracted from the source data all distinct combinations of sentence ID and word position. We excluded sentence-initial and -final RTs from this set since those tend to be substantially slower than the remaining RTs. We then randomly sampled n_{items} with replacement from that set of combinations. These n_{items} were given BATA item IDs from 1 to n_{items}. Next, we extracted all distinct subjects from the source data that had at least one valid RT (> 0) for the sampled item IDs. Then $n_{subjects}$ subjects were sampled with replacement from this set of all subjects. These $n_{subjects}$ were given BATA subject IDs from 1 to $n_{subjects}$. The source data RTs for each combination of BATA item and subject IDs were then read in from the source data (based on the original source data sentence ID, word position, and subject ID) to form the BATA. Responses without valid RTs were dropped.

The RTs for each observation in the BATA then had *up to* $n_{items} * n_{subjects}$ RT observations, though—like in a real experiment—some of the subjects might have missing observations for some of the items.

Post-exclusion BATAs. The post-exclusion BATAs were sampled the same way as the pre-exclusion BATA, except that RTs ≤ 100 ms or ≥ 2000 ms were excluded after sampling. Just like in a real experiment, the post-exclusion BATAs thus had fewer observations, even when they were sampled for the same number of n_{items} and $n_{subjects}$.

Differences in procedure for the NSC data. Unlike the HS13 and F13 data, which were collected in factorially-designed psycholinguistic experiments, the NSC data was collected as a corpus of stories that participants read through. Participants generally only read a subset of the total available stories in the corpus. In order to preserve this structure, each NSC BATA first sampled four stories (with replacement) from all possible stories in the NSC.[18] Subjects without *any* valid responses for more than half of a BATA's sampled stories were excluded from the sampling pool, reducing the amount of missing data and mimicking the way stories were distributed across subjects in the original study. After the stories for a particular BATA were sampled, we sampled $n_{items}/4$ by-word responses from each sampled story (with replacement), keeping the total number of items per subject equal to the equivalent HS18 or F13 BATAs. Then $n_{subjects}$ subjects were sampled (with replacement) from the set of all subjects who had at least one valid RT for at least half of the sampled stories.

SI-2.3 Distributional properties of bootstrap or parametrically generated RTs

A similar picture as presented in the main text emerges for the correlations between the mean, standard deviation, skewness, and kurtosis of RT distributions, which are shown in Figure 25. Unlike RTs generated from the normal distribution, RTs generated under the other parametric assumption resemble natural RT distributions qualitatively in terms of their correlations between the RT means and standard deviations. This again includes the log-transform and log-shifted models (for similar analyses of reaction data from other psychometric tasks, see Baayen & Milin, 2010). However, none of the parametric distributions match the correlation between RT means and skewness or kurtosis found in bootstrapped natural RTs for the pre-exclusion data.

For post-exclusion RTs, differences between source data sets decrease (Figure 25B). Differences between bootstrapped and parametrically generated RTs also are reduced but do persist. For example, while the correlations between moments for the bootstrapped RTs changes noticeably based on the source data, correlations between moments of the parametrically generated data were quite similar across source data.

[18] We chose to sample four stories per NSC BATA to facilitate matching the overall sample sizes of the HS18 and F13 BATAs. For example, sampling 8 stories would yield only a single item per story for the 8x8 BATAs, and 2 stories would have been less similar to the original 10-story data set. Auxiliary Study 2c (SI-6) presents additional simulations that sample 2 or 8 stories per BATA, which replicates the results of Study 2.

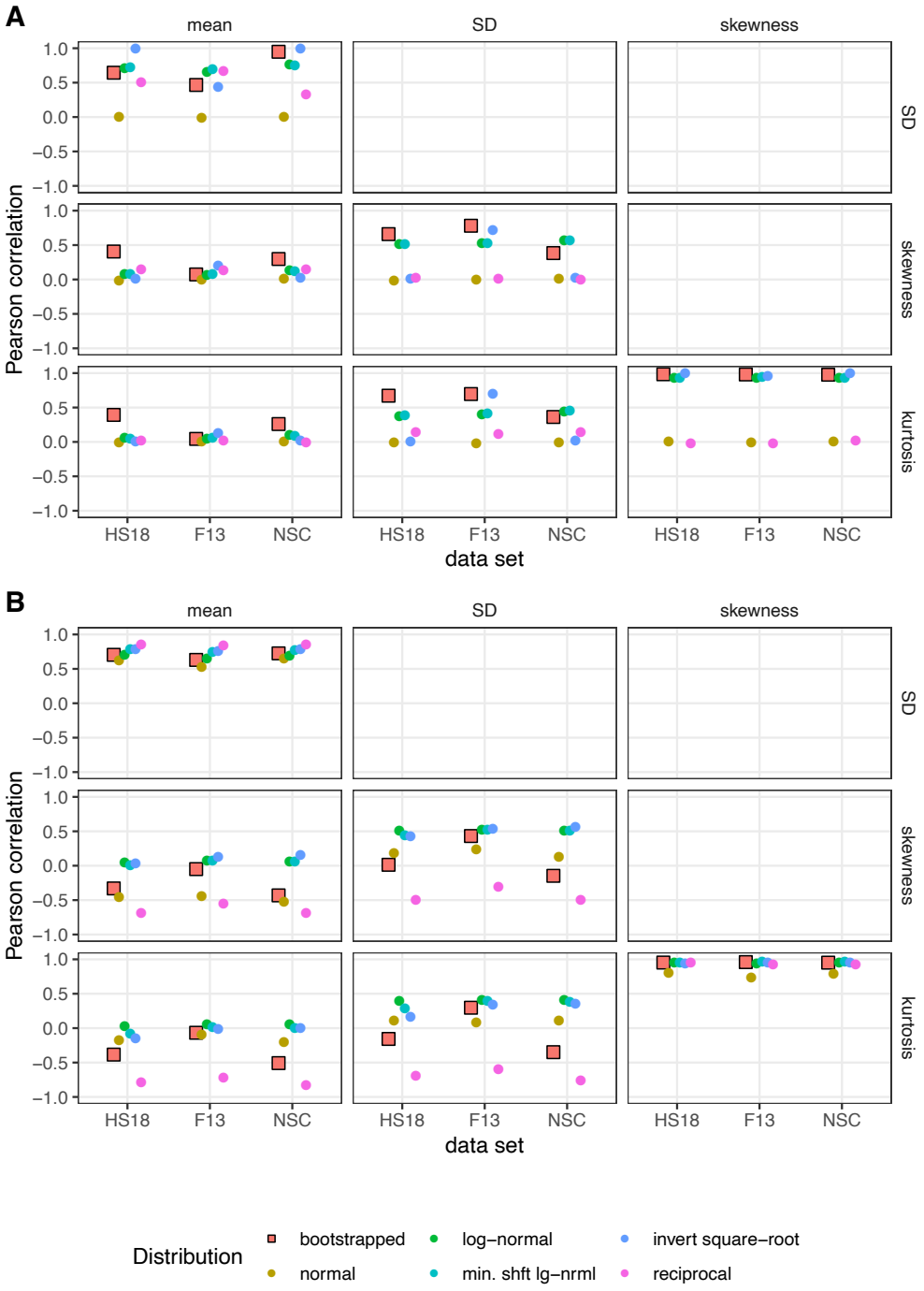

Figure 25. Pearson correlations between the first four statistical moments of the Study 1 RT distributions that are bootstrapped or parametrically generated under various assumptions. The Pearson correlations between statistical moments for the Study 1 BATAs (10,000 for each distribution and data set). The bootstrapped RTs are indicated via outlined rectangles. 95% confidence intervals are shown but too small to be visible. ***Panel A:*** *The correlations for the data generated from the pre-exclusion source data.* ***Panel B:*** *The correlations for the data generated from the post-exclusion source data, after excluding RTs ≤ 100 ms or ≥ 2000 ms.*

SI-3 Study 2

Section SI-3.1 provides additional information on the hierarchical bootstrap procedure for Study 2, depending on the data preparation step. Section SI-3.2 provides additional information on the computational implementation of the analyses we conducted, including how we determined whether an analysis detected an effect for any given BATA. Sections SI-3.3 and SI-3.4 summarize the rate of convergence and singular fits, respectively.[19] Sections SI-3.5 to SI-3.7 provide additional plots and analyses of the Type I error rates, power, and Type I error-corrected power.

SI-3.1 Additional information on the hierarchical bootstrap procedure

We first describe the approach for the HS18 and F13 data, and then describe procedural differences for the NSC data.

Pre-exclusion and post-exclusion BATAs. The approach for the pre- and post-exclusion BATAs was identical to that described for Study 1, except that we added (zero or non-zero) effects to the RTs after sampling. First, we randomly assigned half of these sampled subjects in each BATA to condition A, and the other half to condition B. Then, half of the effect size was subtracted from the RTs of subjects in condition A, and half of the effect size was added to the RTs of subjects in condition B.

Residualized BATAs. The residualized BATAs were sampled the same way as the post-exclusion BATAs. Additionally, we sampled $n_{items} * 15$ word RTs that would serve as "fillers", making the filler-to-critical-item ratio 15:1. While psycholinguistic studies often include only 2 to 3 times as many filler sentences as critical target items, each of the filler sentences tends to contain 5-10 words, so that the ratio of filler *words* to the critical region of the target items is typically much larger than 2:1 or 3:1. Auxiliary Study 2b considers additional filler ratios. Those simulations replicate the results of Study 2.

Any effects that were added to RTs were only added to the sampled *target* items, not the sampled fillers of a BATA. We then fit a linear mixed model to the sampled filler word RTs of the BATA, predicting RTs from word length using the *lmer* function of the *lme4* package (Bates, Mächler, et al., 2015) in R (R Core Team, 2022).[20] This fitted model was used to residualize the RTs of the BATA's target items. The resulting residual (raw or log) RTs of those target items were submitted to the LMM analyses of the effect.

Three-word region BATAs. We first determined all possible regions with three adjacent words (again excluding the first and last word of each sentence). Then n_{items} three-word regions were sampled with replacement in the source data. Residualization was applied *before* the RTs in each region were averaged in a single RT.

Differences in procedure for the NSC data. Unlike the between-subject design of the HS18/F13 BATAs, effects for the NSC BATAs were added between items. For each story, half of the sampled items were assigned to condition A, and the rest were assigned to condition B. As in Study 1, BATAs for Study 2 were sampled from only 4 NSC stories. Auxiliary Study 2c presents additional simulations that sample from fewer or more NSC stories. These additional simulations replicate the results of Study 2.

SI-3.2 Computational implementation of bootstrap and LMM fitting

For every study, sampling and analyses were distributed on a remote compute cluster of 60 machines, with the number of cores for each machine ranging from two to 24. We coordinated the work on the cluster using the *future* (Bengtsson, 2019) and *furrr* (Vaughan & Dancho, 2018) R packages. We ran the different data sets and data preparation conditions at different times, and although the exact number of machines and

---

[19] Singular fits occur when the variance-covariance matrix is non-invertible (e.g., because variance parameters in the model are computationally indistinguishable from zero). While singular fits do not necessarily indicate a failure to converge, they can be indicative of over-fitting and can support misleading statistical inferences (for discussion, see Bates, Kliegl, et al., 2015).

[20] The following formula was used RT ~ 1 + Word.Length + (1 + Word.Length | Subject), where RT was the raw or log-transformed RTs.

cores per machine used varied, the total *compute* time for Study 2 was approximately three days, averaging 30 machines with six cores each. These compute times do not take into account debugging and coordination of the computations. In total, Study 2 took more than 2 months to complete.

Analyses were run with the default *lmer* settings (e.g., default tolerance, maximum iteration, etc.), but with the *BOBYQA* optimizer, which, while generally slower than the default *nloptwrap* optimizer, has a higher rate of convergence. All models were fit to maximize the restricted likelihood (REML). Different choices of optimizers and control settings might affect convergence rates and the rates of singular fits, but are unlikely to change the Type I and power results that constitute the focus of the present studies. Relatedly, we note that we did not standardize RTs or log-RTs. This, too, might affect convergence rates, but should not affect Type I and power results. Since convergence rates were >99% in all simulations, these decisions cannot confound our results.

P-values were derived via Satterthwaite approximation for degrees of freedom (Satterthwaite, 1941) from the *lmerTest* R package (Kuznetsova et al., 2017).

SI-3.3 Convergence failures

Many researchers determine the random effect structure of their analyses based on convergence (following influential but potentially problematic advice, Barr, Levy, Scheepers, & Tily, 2013; for discussion, see Bates et al., 2015). We thus report convergence rates for our analyses.

Convergence rates were high across all BATA conditions, data sets, and simulations. The highest number of convergence failures per simulation was 22 out of 10,000 (.2%). As shown in Figure 26 for the HS18 data, effect size did not meaningfully affect convergence rates. Figure 27 therefore averages over effect sizes for the F13 and NSC data. The untransformed analysis approach resulted in more convergence failures compared to the log-transformed analysis (see Figure 26 and Figure 27).

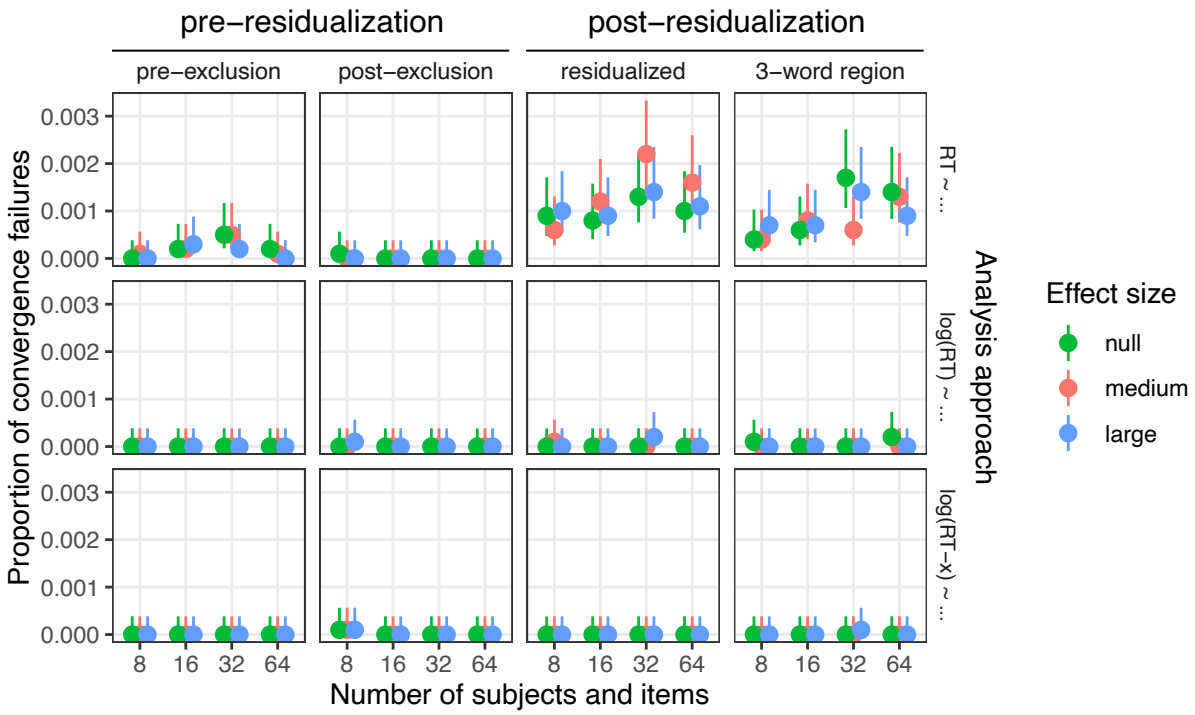

Figure 26. The rate of convergence failures in all simulation conditions for the HS18 BATAs in Study 2. Overall, converge failures were rare—less than 1%. Lines show 95% binomial confidence intervals. See SI-3 for identical figures for the F13 and NSC data.

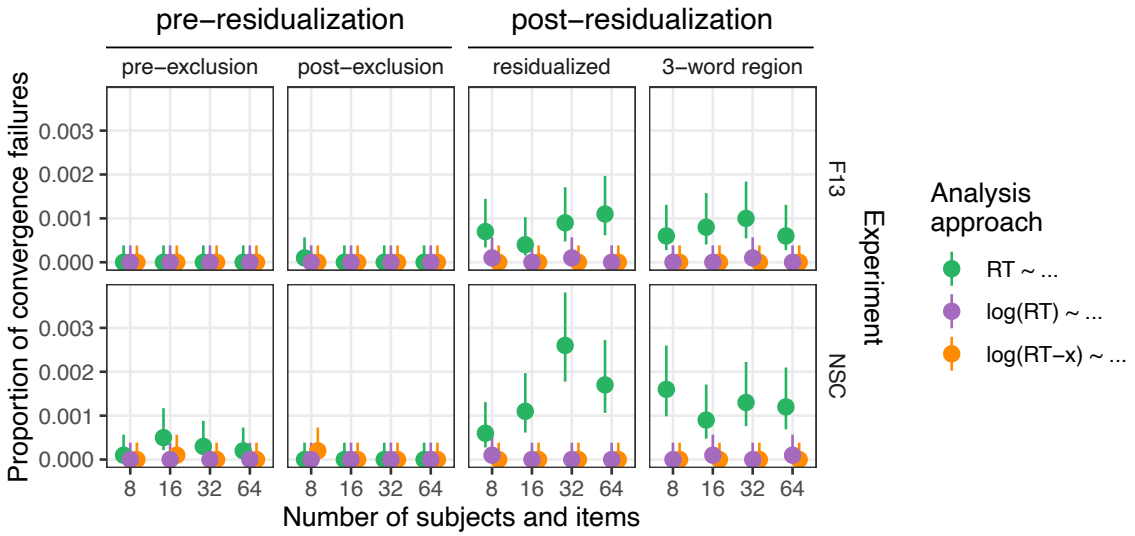

Figure 27. Rate of convergence failures by sample size and data preparation for the F13 and NSC BATAs in Study 2, averaged across effect sizes. Overall, converge failures were rare—less than 1%.

To compare the rate of convergence failure between analysis approaches, we used a logistic regression to predict the probability of convergence from the full factorial of sample size, effect size (null, medium, or large), data preparation condition, and analysis approach. We conducted separate logistic regressions for each of the three source data sets. Specifically, we used χ^2-tests over the difference in deviance to simplify the full logistic regression (containing the full factorial of sample size, effect size, data preparation condition, and analysis approach). At each step, we asked whether removal of the highest-order terms that included analysis approach resulted in a model with significantly worse deviance. We did this for every level of interaction, comparing the model against the next simpler model (by again removing the highest-order terms containing analysis approach; e.g., all three-way interactions including the analysis approach, all two-way interactions including the analysis approach, etc.).[21] This was repeated until all analysis approach terms were removed. If any of these model comparisons were significant, we report the significant interactions with analysis approach.

HS18. The log-transformed analysis approach had lower rates of convergence failures than the untransformed approach for 34 of the 35 comparisons that were neither at ceiling nor at floor (of which there were 13). Stepwise model comparison found that analysis approach had a significant effect on rate of convergence failure with the HS18 simulations (Table 2).

Table 2: Results of nested model comparison of convergence failures in HS18 BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	18	3.94	1.00e+00	
2	21	4.53	1.00e+00	
3	8	31.19	1.30e-04	***
4	1	208.35	3.14e-47	***

[21] When a logistic regression resulted in a singularity (e.g., because *all* analyses in a set of BATAs converged), we determined the minimally simplified logistic regression that avoided this singularity.

After removing the pre-residualized data preparation steps to avoid singularities, the full interaction model found that the log-transformed analysis had significantly fewer convergence failures than the untransformed approach (p<0.001). None of the interactions with analysis approach were significant.

F13. The log-transformed analysis approach had lower rates of convergence failures than the untransformed approach for 26 of the 27 comparisons that were neither at ceiling nor at floor (of which there were 21). Stepwise model comparison found that analysis approach had a significant effect on rates of convergence failure with the F13 simulations (Table 3).

Table 3: Results of nested model comparison of convergence failures in F13 BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	p	
1	18	3.19	1.00e+00	
2	21	7.92	9.95e-01	
3	8	33.59	4.83e-05	***
4	1	135.89	2.10e-31	***

After removing the pre-residualized data preparation steps to avoid singularities, the full interaction model found that the log-transformed analysis had significantly fewer convergence failures than the untransformed approach (p<0.001).

NSC. The log-transformed analysis approach had lower rates of convergence failures than the untransformed approach for 37 of the 37 comparisons that were neither at ceiling nor at floor (of which there were 11). Stepwise model comparison found that analysis approach had a significant effect on rates of convergence failure with the NSC simulations (Table 4).

Table 4: Results of nested model comparison of convergence failures in NSC BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	p	
1	18	4.60	9.99e-01	
2	21	4.95	1.00e+00	
3	8	29.99	2.12e-04	***
4	1	271.06	6.68e-61	***

After removing the pre-residualized data preparation steps to avoid singularities, the full interaction model found that the log-transformed analysis had significantly fewer convergence failures than the untransformed approach (p<0.001), and that the full four-way interaction was marginally significant (p=0.095).

SI-3.4 Singular fits

While singular fits are not *necessarily* problematic, singular fits can indicate over-fitting and other issues. We thus also report the rate of singular fits. The rate of singular fits was substantially higher than convergence failures. As shown in Figure 28 for HS18, the rate of singular fits increased substantially for smaller sample sizes. Paralleling convergence failures, effect size did not meaningfully affect the rate of singular fits. Figure 29 therefore averages across effect sizes for F13 and NSC. The untransformed analysis approach resulted in more singular fits, compared to the log-transformed analysis (see Figure 28 and Figure 29).

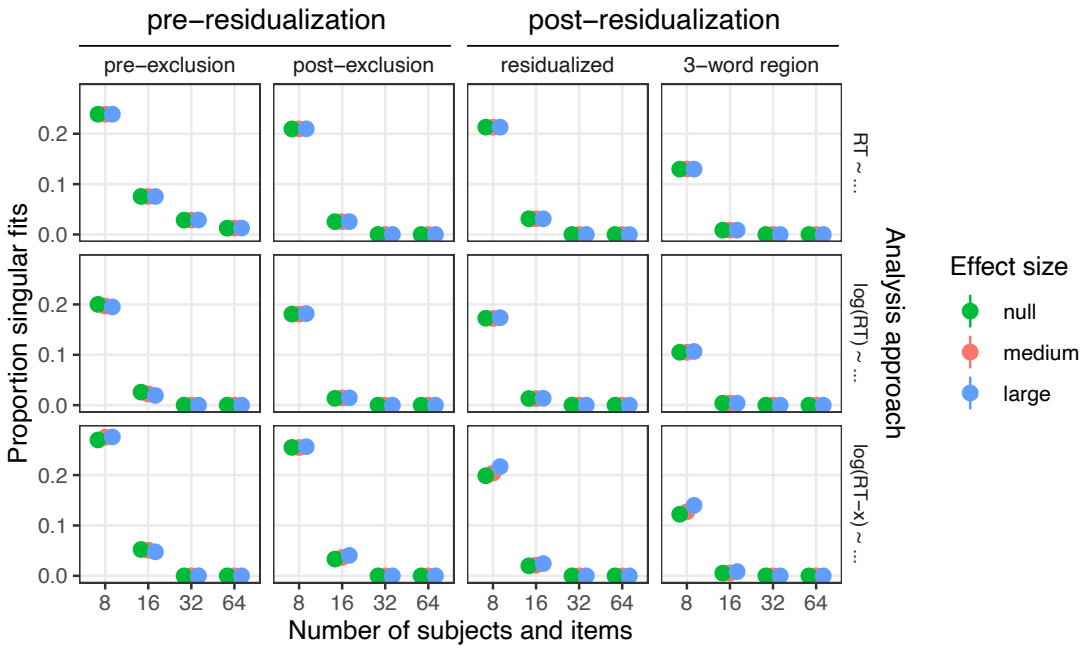

Figure 28. Rate of singular fits in all simulation conditions for the HS18 BATAs in Study 2. As expected, the greatest determiner of singular fits is sample size. Effect size and the presence/absence of an effect have little impact on this rate. Vertical lines—obscured in many cases because they were small—show 95% binomial confidence intervals.

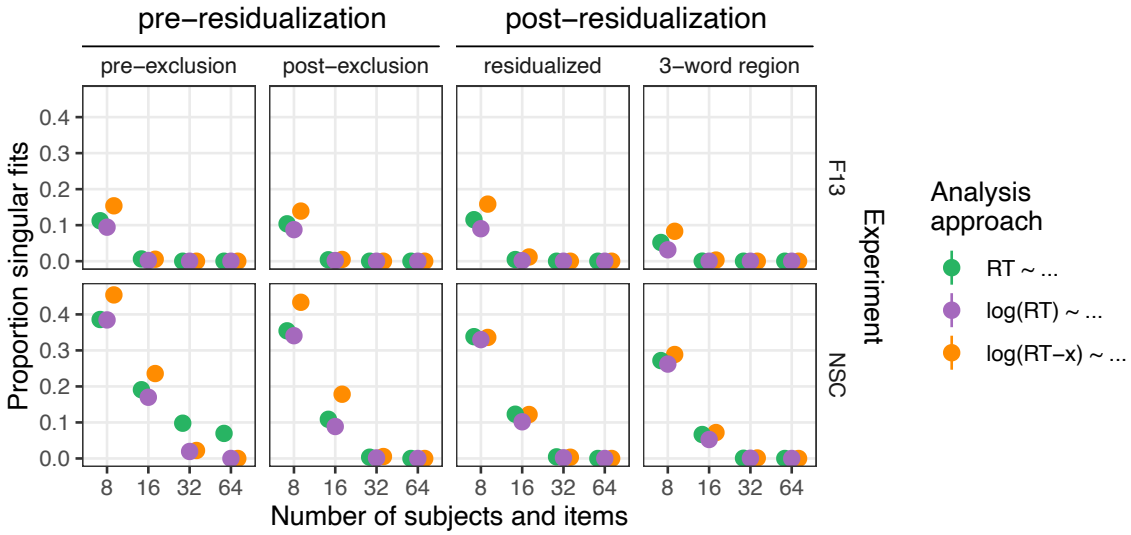

Figure 29. Rate of singular fits by sample size and data preparation for F13 and NSC BATAs in Study 2, averaged across effect sizes. For additional details, see Figure 28.

We analyzed singular fit rates following the same approach as for convergence failures.

HS18. The log-transformed analysis approach had lower rates of singular fit than the untransformed approach for 36 of the 36 comparisons that were neither at ceiling nor at floor. Stepwise model comparison found that analysis approach had a significant effect on rates of singular fit with the HS18 simulations (Table 5).

Table 5: Results of nested model comparison of singular fit rates in HS18 BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	18	4.23	1.00e+00	
2	21	186.07	1.95e-28	***
3	8	1,795.05	0.00e+00	***
4	1	1,477.62	2.96e-323	***

After removing the 64x64 and 32x32 sample size simulations to avoid singularities, the full interaction model found that the log-transformed analysis had significantly fewer singular fits than the untransformed approach (p<0.001). It also had a number of significant interactions with data preparation.

F13. the log-transformed analysis approach had lower rates of singular fit than the untransformed approach for 24 of the 24 comparisons that were neither at ceiling nor at floor. Stepwise model comparison found that analysis approach had a significant effect on rates of singular fit with the F13 simulations (Table 6).

Table 6: Results of nested model comparison of singular fit rates in F13 BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	18	0.621	1.00e+00	
2	21	2.569	1.00e+00	
3	8	131.422	1.43e-24	***
4	1	370.965	1.15e-82	***

After removing the 16x16, 32x32, and 64x64 sample size simulations to avoid singularities, the full interaction model found that the log-transformed analysis had significantly fewer singular fits than the untransformed approach (p<0.001). This effect interacted significantly with data preparation condition, so that the advantage of the log-transformed analysis was reduced for the pre-exclusion (p<0.001) and post-exclusion condition (p<0.001).

NSC. The log-transformed analysis approach had lower rates of singular fit than the untransformed approach for 39 of the 39 comparisons that were neither at ceiling nor at floor. Stepwise model comparison found that analysis approach had a significant effect on rates of singular fit with the NSC simulations (Table 7).

Table 7: Results of nested model comparison of singular fit rates in NSC BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	18	24.7	1.34e-01	
2	21	318.4	5.51e-55	***
3	8	4,571.0	0.00e+00	***
4	1	1,034.2	6.57e-227	***

After removing the 32x32 and 64x64 sample size simulations to avoid singularities, the full interaction model found that the log-transformed analysis had significantly fewer singular fits than the untransformed approach (p<0.001). This effect interacted significantly with data preparation condition, so that the advantage of the log-transformed analysis was stronger for the pre-exclusion condition (p<0.001) but reduced for the residualized condition (p=0.053). It also significantly interacted with effect size, being

stronger for the null effect size (p=0.03) and weaker for the medium effect size (p=0.013), with the higher-order interactions of these two factors also being significant.

SI-3.5 Type I errors

We analyzed Type I error rates following the same procedure used for convergence rates and singular fits.

HS18. For the HS18 simulations, the log-transformed analysis approach had higher Type I error rates than the untransformed approach for 14 of the 16 comparisons that were neither at ceiling nor at floor. The log-transformed analysis approach had higher Type I error rates on average, with a mean difference of 0.0378 in log-odds, which corresponds to a difference of 0.18% in Type I error when centered around 5%. Stepwise model comparison found that analysis approach had a significant effect on Type I error rates with the HS18 simulations (Table 8).

Table 8: Results of nested model comparison of Type I error rates in HS18 BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	p	
1	9	2.09	0.9899	
2	6	2.52	0.8659	
3	1	5.39	0.0203	*

The full interaction model found that the log-transformed analysis had significantly higher Type I error rates than the untransformed approach (p=0.029), but there were no significant interactions.

Both analysis approaches were significantly conservative. The untransformed analysis approach had an average Type I error rate of 0.0436, significantly lower than 0.05 (p<0.001). The log-transformed analysis approach had an average Type I error rate of 0.0452—about 26% closer to, but still significantly smaller than, 0.05 (p<0.001).

Table 9: The summary of the full interaction model for the Type I HS18 data in Study 2.

term	estimate	std.error	statistic	p.value	
(Intercept)	-3.07	0.00863	-356.	< 0.001	***
SampSize-8	0.0126	0.0149	0.850	0.395	
SampSize-16	-0.0122	0.0150	-0.814	0.416	
SampSize-32	0.00789	0.0149	0.529	0.597	
DataPrep-pre-exclusion	0.104	0.0144	7.21	5.7e-13	***
DataPrep-post-exclusion	0.116	0.0144	8.05	8.3e-16	***
DataPrep-residualized	0.0243	0.0148	1.64	0.100	
Anlys-stndrd	-0.0189	0.00863	-2.19	0.029	*
SampSize-8:DataPrep-pre-exclusion	0.00192	0.0249	0.0771	0.939	
SampSize-16:DataPrep-pre-exclusion	0.00851	0.0251	0.339	0.734	
SampSize-32:DataPrep-pre-exclusion	0.0245	0.0248	0.984	0.325	
SampSize-8:DataPrep-post-exclusion	-0.0528	0.0251	-2.10	0.035	*
SampSize-16:DataPrep-post-exclusion	0.0274	0.0249	1.10	0.271	
SampSize-32:DataPrep-post-exclusion	-0.0282	0.0250	-1.13	0.260	
SampSize-8:DataPrep-residualized	0.0299	0.0254	1.18	0.239	
SampSize-16:DataPrep-residualized	0.0313	0.0256	1.23	0.220	
SampSize-32:DataPrep-residualized	0.00362	0.0255	0.142	0.887	
SampSize-8:Anlys-stndrd	-0.00909	0.0149	-0.611	0.541	
SampSize-16:Anlys-stndrd	-0.00333	0.0150	-0.222	0.825	
SampSize-32:Anlys-stndrd	0.00191	0.0149	0.128	0.898	
DataPrep-pre-exclusion:Anlys-stndrd	-0.0155	0.0144	-1.07	0.284	
DataPrep-post-exclusion:Anlys-stndrd	-0.00567	0.0144	-0.394	0.694	

term	estimate	std.error	statistic	p.value
DataPrep-residualized:Anlys-stndrd	0.00703	0.0148	0.475	0.634
SampSize-8:DataPrep-pre-exclusion:Anlys-stndrd	-0.00203	0.0249	-0.0816	0.935
SampSize-16:DataPrep-pre-exclusion:Anlys-stndrd	-0.0138	0.0251	-0.551	0.582
SampSize-32:DataPrep-pre-exclusion:Anlys-stndrd	0.0109	0.0248	0.437	0.662
SampSize-8:DataPrep-post-exclusion:Anlys-stndrd	-0.00920	0.0251	-0.367	0.714
SampSize-16:DataPrep-post-exclusion:Anlys-stndrd	0.00905	0.0249	0.363	0.716
SampSize-32:DataPrep-post-exclusion:Anlys-stndrd	-0.0108	0.0250	-0.434	0.665
SampSize-8:DataPrep-residualized:Anlys-stndrd	0.00145	0.0254	0.0572	0.954
SampSize-16:DataPrep-residualized:Anlys-stndrd	0.0122	0.0256	0.478	0.633
SampSize-32:DataPrep-residualized:Anlys-stndrd	0.0153	0.0255	0.599	0.549

F13. The log-transformed analysis approach had higher Type I error rates than the untransformed approach for 11 of the 16 comparisons that were neither at ceiling nor at floor. The log-transformed analysis approach had higher Type I error rates on average, with a mean difference of 0.0171 in log-odds, which corresponds to a difference of 0.081% Type I error when centered around the targeted error of 5%. Stepwise model comparison found that analysis approach did not have a significant effect on Type I error rates with the F13 simulations (Table 10).

Table 10: Results of nested model comparison of Type I error rates in F13 BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*
1	9	1.82	0.994
2	6	1.53	0.957
3	1	1.06	0.304

Although the omnibus tests did not reveal a significant effect of analysis approach for the F13 data on Type I error rates, a post-hoc look at the full interaction model showed that numerically (p=0.32), the log-transformed analysis approach had higher Type I error rates.

Both analysis approaches were, on average, conservative. The untransformed analysis approach had a Type I error rate of 0.0449, significantly below 0.05 (p<0.001). The log-transformed analysis approach had an average Type I error rate of 0.0456—about 15% closer to, but still significantly smaller than, 0.05 (p<0.001).

Table 11: The summary of the full interaction model for the Type I F13 data in Study 2.

term	estimate	std.error	statistic	p.value	
(Intercept)	-3.05	0.00858	-356.	<2e-16	***
SampSize-8	0.00492	0.0149	0.331	0.74	
SampSize-16	-0.0166	0.0149	-1.11	0.27	
SampSize-32	-0.0188	0.0149	-1.26	0.21	
DataPrep-pre-exclusion	0.153	0.0141	10.8	<2e-16	***
DataPrep-post-exclusion	0.138	0.0142	9.69	<2e-16	***
DataPrep-residualized	0.00370	0.0148	0.251	0.80	
Anlys-stndrd	-0.00857	0.00858	-0.999	0.32	
SampSize-8:DataPrep-pre-exclusion	0.0249	0.0243	1.02	0.31	
SampSize-16:DataPrep-pre-exclusion	-0.0255	0.0247	-1.03	0.30	
SampSize-32:DataPrep-pre-exclusion	-0.00768	0.0247	-0.311	0.76	
SampSize-8:DataPrep-post-exclusion	0.0140	0.0245	0.573	0.57	
SampSize-16:DataPrep-post-exclusion	0.0165	0.0246	0.672	0.50	
SampSize-32:DataPrep-post-exclusion	0.00105	0.0247	0.0425	0.97	
SampSize-8:DataPrep-residualized	0.00128	0.0255	0.0500	0.96	

term	estimate	std.error	statistic	p.value
SampSize-16:DataPrep-residualized	-0.0001	0.0257	-0.0048	1.00
SampSize-32:DataPrep-residualized	0.00876	0.0257	0.341	0.73
SampSize-8:Anlys-stndrd	0.00417	0.0149	0.281	0.78
SampSize-16:Anlys-stndrd	0.0104	0.0149	0.698	0.48
SampSize-32:Anlys-stndrd	-0.0131	0.0149	-0.878	0.38
DataPrep-pre-exclusion:Anlys-stndrd	-0.00100	0.0141	-0.0709	0.94
DataPrep-post-exclusion:Anlys-stndrd	-0.00650	0.0142	-0.458	0.65
DataPrep-residualized:Anlys-stndrd	0.0101	0.0148	0.682	0.49
SampSize-8:DataPrep-pre-exclusion:Anlys-stndrd	-0.0123	0.0243	-0.505	0.61
SampSize-16:DataPrep-pre-exclusion:Anlys-stndrd	-0.00819	0.0247	-0.331	0.74
SampSize-32:DataPrep-pre-exclusion:Anlys-stndrd	0.0123	0.0247	0.501	0.62
SampSize-8:DataPrep-post-exclusion:Anlys-stndrd	-0.0152	0.0245	-0.622	0.53
SampSize-16:DataPrep-post-exclusion:Anlys-stndrd	0.00976	0.0246	0.396	0.69
SampSize-32:DataPrep-post-exclusion:Anlys-stndrd	0.00532	0.0247	0.215	0.83
SampSize-8:DataPrep-residualized:Anlys-stndrd	0.0188	0.0255	0.736	0.46
SampSize-16:DataPrep-residualized:Anlys-stndrd	-0.00091	0.0257	-0.0354	0.97
SampSize-32:DataPrep-residualized:Anlys-stndrd	-0.0239	0.0257	-0.931	0.35

NSC. The log-transformed analysis approach had higher Type I error rates than the untransformed approach for 10 of the 16 comparisons that were not both at ceiling or floor. The log-transformed analysis approach had higher Type I error rates on average, with a mean difference of 0.0676 in log-odds, which corresponds to a difference of 0.32% Type I error when centered around the targeted 5%. Stepwise model comparison found that analysis approach had a significant effect on Type I error rates with the NSC simulations (Table 12).

Table 12: Results of nested model comparison of Type I error rates in NSC BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	9	23.1	6.00e-03	**
2	6	90.3	2.61e-17	***
3	1	30.7	3.06e-08	***

The full interaction model found that the log-transformed analysis had significantly higher Type I error rates than the untransformed approach ($p<0.001$). There were, however, significant interactions with data preparation condition, with the pre-exclusion condition having less of an effect ($p<0.001$), and the post-exclusion condition having a greater effect ($p=0.03$).

Both analysis approaches were, on average, conservative—even more so than for the other two source data. The untransformed analysis approach had a Type I error rate of 0.0339—about 15% closer to, but still significantly smaller than, 0.05 ($p<0.001$).

Table 13: The summary of the full interaction model for the Type I NSC data in Study 2.

term	estimate	std.error	statistic	p.value	
(Intercept)	-3.32	0.00967	-343.	< 0.001	***
SampSize-8	-0.127	0.0174	-7.29	3.1e-13	***
SampSize-16	-0.0107	0.0168	-0.638	0.52327	
SampSize-32	0.0731	0.0164	4.46	8.0e-06	***
DataPrep-pre-exclusion	-0.162	0.0176	-9.20	< 0.001	***
DataPrep-post-exclusion	0.0219	0.0166	1.32	0.18686	
DataPrep-residualized	0.0908	0.0163	5.58	2.4e-08	***
Anlys-stndrd	-0.0338	0.00967	-3.50	0.00047	***
SampSize-8:DataPrep-pre-exclusion	0.0635	0.0313	2.03	0.04205	*

term	estimate	std.error	statistic	p.value	
SampSize-16:DataPrep-pre-exclusion	-0.0111	0.0307	-0.363	0.71689	
SampSize-32:DataPrep-pre-exclusion	-0.0682	0.0303	-2.25	0.02446	*
SampSize-8:DataPrep-post-exclusion	-0.00311	0.0299	-0.104	0.91706	
SampSize-16:DataPrep-post-exclusion	0.000348	0.0288	0.0121	0.99035	
SampSize-32:DataPrep-post-exclusion	-0.0137	0.0282	-0.487	0.62636	
SampSize-8:DataPrep-residualized	-0.0499	0.0296	-1.69	0.09131	
SampSize-16:DataPrep-residualized	-0.0303	0.0284	-1.07	0.28523	
SampSize-32:DataPrep-residualized	0.0651	0.0272	2.40	0.01660	*
SampSize-8:Anlys-stndrd	0.00778	0.0174	0.447	0.65456	
SampSize-16:Anlys-stndrd	-0.0166	0.0168	-0.989	0.32282	
SampSize-32:Anlys-stndrd	0.0113	0.0164	0.689	0.49062	
DataPrep-pre-exclusion:Anlys-stndrd	-0.101	0.0176	-5.74	9.7e-09	***
DataPrep-post-exclusion:Anlys-stndrd	0.0361	0.0166	2.18	0.02952	*
DataPrep-residualized:Anlys-stndrd	0.0184	0.0163	1.13	0.25874	
SampSize-8:DataPrep-pre-exclusion:Anlys-stndrd	0.0525	0.0313	1.68	0.09326	
SampSize-16:DataPrep-pre-exclusion:Anlys-stndrd	0.0394	0.0307	1.29	0.19848	
SampSize-32:DataPrep-pre-exclusion:Anlys-stndrd	-0.0231	0.0303	-0.762	0.44621	
SampSize-8:DataPrep-post-exclusion:Anlys-stndrd	-0.0460	0.0299	-1.54	0.12343	
SampSize-16:DataPrep-post-exclusion:Anlys-stndrd	-0.00467	0.0288	-0.162	0.87125	
SampSize-32:DataPrep-post-exclusion:Anlys-stndrd	0.0344	0.0282	1.22	0.22226	
SampSize-8:DataPrep-residualized:Anlys-stndrd	-0.0251	0.0296	-0.849	0.39563	
SampSize-16:DataPrep-residualized:Anlys-stndrd	-0.0184	0.0284	-0.647	0.51736	
SampSize-32:DataPrep-residualized:Anlys-stndrd	0.0133	0.0272	0.488	0.62577	

SI-3.6 Power (uncorrected)

HS18. The log-transformed analysis approach had higher power than the untransformed approach for 30 of the 30 comparisons that were neither at ceiling nor at floor (of which there were 2). The log-transformed analysis approach had higher power on average, with a mean difference of 1.36 in log-odds (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of, e.g., 91% vs. 69% power.

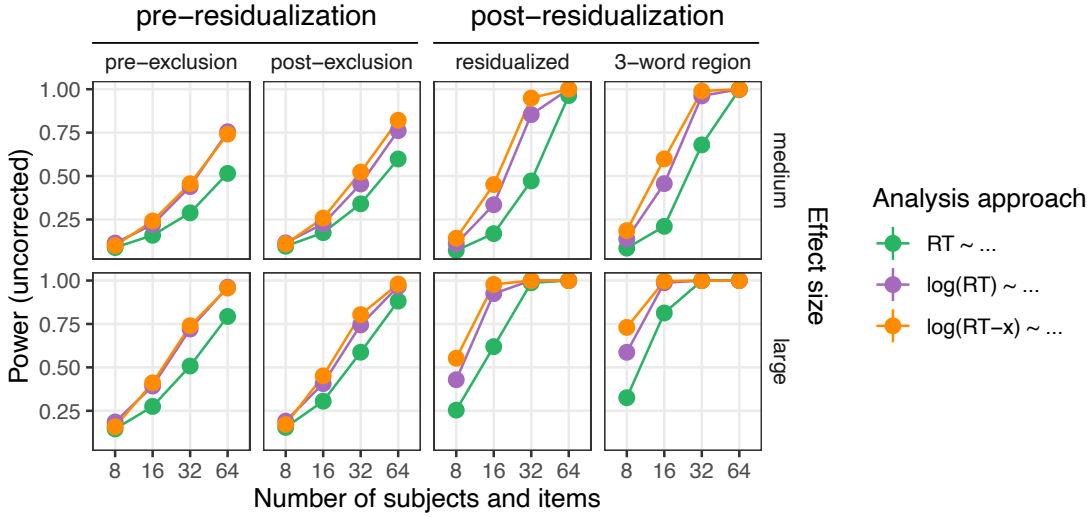

Figure 30: Power (uncorrected) for both analysis approaches on the HS18 data. Effect sizes have been collapsed into two categories: 'medium' (15 and 56 ms) and 'large' (35 and 80 ms).

F13. The log-transformed analysis approach had higher power than the untransformed approach for 28 of the 28 comparisons that were neither at ceiling nor at floor (of which there were 4). The log-transformed

analysis approach had higher power on average, with a mean difference of 1.02 in log-odds (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of, e.g., 88% vs. 72% power.

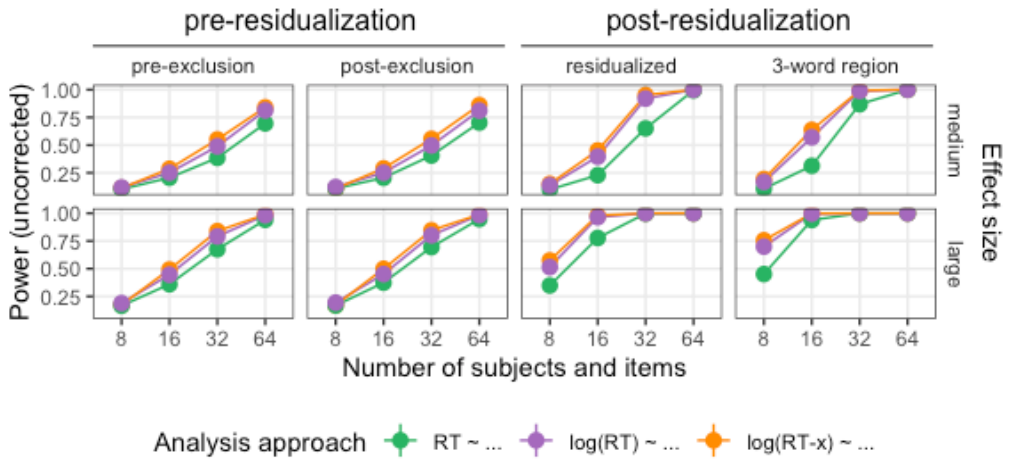

Figure 31: Power (uncorrected) for both analysis approaches on the F13 data. Effect sizes have been collapsed into two categories: 'medium' (15 and 56 ms) and 'large' (35 and 80 ms).

NSC. The log-transformed analysis approach had higher power than the untransformed approach for 31 of the 31 comparisons that were neither at ceiling nor at floor (of which there were 1). The log-transformed analysis approach had, on average, 2.02 in log-odds more power (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of, e.g., 96% vs. 64% power.

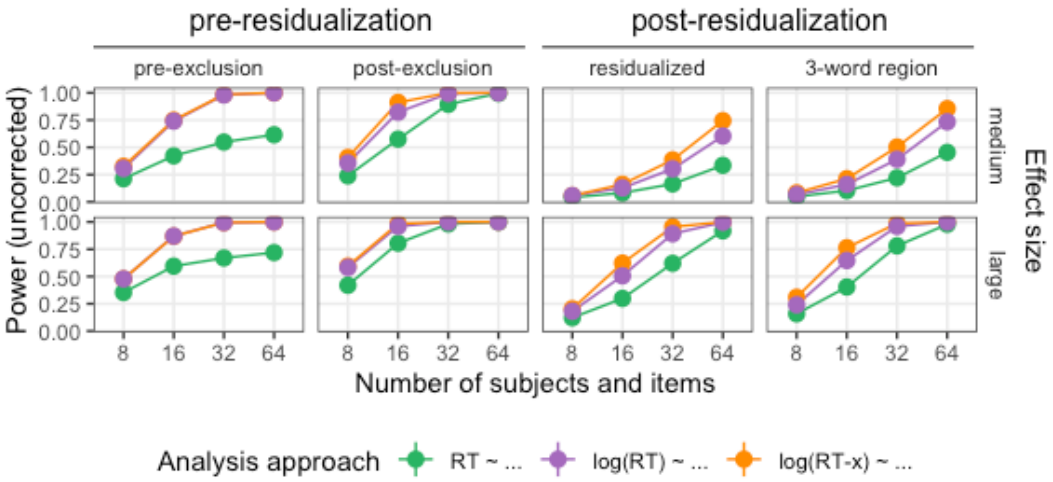

Figure 32: Power (uncorrected) for both analysis approaches on the NSC data. Effect sizes have been collapsed into two categories: 'medium' (15 and 56 ms) and 'large' (35 and 80 ms).

SI-3.7 Power (corrected for Type I error rate)

HS18. The log-transformed analysis approach had higher Type I-corrected power than the untransformed approach for 30 of the 30 comparisons that were neither at ceiling nor at floor (of which there were 2). The log-transformed analysis approach had, on average, 1.32 in log-odds higher Type I-corrected power (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of, e.g., 90.5% vs. 69.5% power.

The nested model comparison found that analysis approach had a significant effect on Type I-corrected power with the HS18 simulations (Table 14).

Table 14: Results of nested model comparison of Type I-corrected power in HS18 BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	9	157	3.17e-29	***
2	15	1,295	5.04e-267	***
3	7	5,180	0.00e+00	***
4	1	15,204	0.00e+00	***

After removing the 32x32 and 64x64 sample size simulations to avoid singularities, the full interaction model found that the log-transformed analysis had significantly better power than the untransformed approach ($p<0.001$). Every term in the follow model, including all the interactions, was significant ($ps<0.0001$).

Table 15: The summary of the full interaction model for the HS18 Type I-corrected power in Study 2.

term	estimate	std.error	statistic	p.value	
(Intercept)	-0.804	0.00586	-137.	< 0.001	***
SampSize-8	-0.704	0.00586	-120.	< 0.001	***
DataPrep-pre-exclusion	-0.678	0.00896	-75.7	< 0.001	***
DataPrep-post-exclusion	-0.601	0.00883	-68.0	< 0.001	***
DataPrep-residualized	0.195	0.00916	21.3	< 0.001	***
Effect-medium	-0.786	0.00586	-134.	< 0.001	***
Anlys-stndrd	-0.381	0.00586	-64.9	< 0.001	***
SampSize-8:DataPrep-pre-exclusion	0.289	0.00896	32.3	< 0.001	***
SampSize-8:DataPrep-post-exclusion	0.292	0.00883	33.1	< 0.001	***
SampSize-8:DataPrep-residualized	-0.127	0.00916	-13.9	< 0.001	***
SampSize-8:Effect-medium	0.210	0.00586	35.9	< 0.001	***
DataPrep-pre-exclusion:Effect-medium	0.454	0.00896	50.7	< 0.001	***
DataPrep-post-exclusion:Effect-medium	0.446	0.00883	50.5	< 0.001	***
DataPrep-residualized:Effect-medium	-0.279	0.00916	-30.4	< 0.001	***
SampSize-8:Anlys-stndrd	0.154	0.00586	26.4	< 0.001	***
DataPrep-pre-exclusion:Anlys-stndrd	0.233	0.00896	26.0	< 0.001	***
DataPrep-post-exclusion:Anlys-stndrd	0.251	0.00883	28.4	< 0.001	***
DataPrep-residualized:Anlys-stndrd	-0.139	0.00916	-15.2	< 0.001	***
Effect-medium:Anlys-stndrd	0.137	0.00586	23.3	< 0.001	***
SampSize-8:DataPrep-pre-exclusion:Effect-medium	-0.161	0.00896	-18.0	< 0.001	***
SampSize-8:DataPrep-post-exclusion:Effect-medium	-0.154	0.00883	-17.4	< 0.001	***
SampSize-8:DataPrep-residualized:Effect-medium	0.0451	0.00916	4.92	8.6e-07	***
SampSize-8:DataPrep-pre-exclusion:Anlys-stndrd	-0.102	0.00896	-11.4	< 0.001	***
SampSize-8:DataPrep-post-exclusion:Anlys-stndrd	-0.113	0.00883	-12.8	< 0.001	***
SampSize-8:DataPrep-residualized:Anlys-stndrd	0.0583	0.00916	6.36	2.0e-10	***
SampSize-8:Effect-medium:Anlys-stndrd	-0.0751	0.00586	-12.8	< 0.001	***
DataPrep-pre-exclusion:Effect-medium:Anlys-stndrd	-0.115	0.00896	-12.8	< 0.001	***
DataPrep-post-exclusion:Effect-medium:Anlys-stndrd	-0.114	0.00883	-12.9	< 0.001	***
DataPrep-residualized:Effect-medium:Anlys-stndrd	0.0436	0.00916	4.76	1.9e-06	***
SampSize-8:DataPrep-pre-exclusion:Effect-medium:Anlys-stndrd	0.0606	0.00896	6.77	1.3e-11	***
SampSize-8:DataPrep-post-exclusion:Effect-medium:Anlys-stndrd	0.0649	0.00883	7.35	2.0e-13	***
SampSize-8:DataPrep-residualized:Effect-medium:Anlys-stndrd	-0.0243	0.00916	-2.66	0.0079	**

F13. The log-transformed analysis approach had higher Type I-corrected power than the untransformed approach for 28 of the 28 comparisons that were neither at ceiling nor at floor (of which there were 4). The log-transformed analysis approach had, on average, 1.01 log-odds higher Type I-corrected power (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of, e.g., 88% vs. 72% power.

The nested model comparison found that analysis approach had a significant effect on Type I-corrected power with the F13 simulations (Table 16).

Table 16: Results of nested model comparison of Type I-corrected power in F13 BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	9	141	7.60e-26	***
2	15	867	3.21e-175	***
3	7	3,298	0.00e+00	***
4	1	7,706	0.00e+00	***

After removing the 32x32 and 64x64 sample size simulations to avoid singularities due to ceiling effects, the full interaction model found that the log-transformed analysis had significantly better power than the untransformed approach ($p<0.001$). Every term in the model, including all the interactions, was significant ($ps<0.0001$).

Table 17: The summary of the full interaction model for the F13 Type I-corrected power in Study 2.

term	estimate	std.error	statistic	p.value	
(Intercept)	-0.459	0.00882	-52.1	< 0.001	***
SampSize-8	-0.849	0.00882	-96.2	< 0.001	***
DataPrep-pre-exclusion	-0.891	0.0110	-80.8	< 0.001	***
DataPrep-post-exclusion	-0.863	0.0110	-78.7	< 0.001	***
DataPrep-residualized	0.330	0.0117	28.3	< 0.001	***
Effect-medium	-0.899	0.00882	-102.	< 0.001	***
Anlys-stndrd	-0.359	0.00882	-40.6	< 0.001	***
SampSize-8:DataPrep-pre-exclusion	0.311	0.0110	28.2	< 0.001	***
SampSize-8:DataPrep-post-exclusion	0.339	0.0110	30.9	< 0.001	***
SampSize-8:DataPrep-residualized	-0.0966	0.0117	-8.26	< 0.001	***
SampSize-8:Effect-medium	0.281	0.00882	31.9	< 0.001	***
DataPrep-pre-exclusion:Effect-medium	0.553	0.0110	50.1	< 0.001	***
DataPrep-post-exclusion:Effect-medium	0.550	0.0110	50.1	< 0.001	***
DataPrep-residualized:Effect-medium	-0.312	0.0117	-26.7	< 0.001	***
SampSize-8:Anlys-stndrd	0.170	0.00882	19.3	< 0.001	***
DataPrep-pre-exclusion:Anlys-stndrd	0.259	0.0110	23.5	< 0.001	***
DataPrep-post-exclusion:Anlys-stndrd	0.264	0.0110	24.1	< 0.001	***
DataPrep-residualized:Anlys-stndrd	-0.143	0.0117	-12.2	< 0.001	***
Effect-medium:Anlys-stndrd	0.145	0.00882	16.4	< 0.001	***
SampSize-8:DataPrep-pre-exclusion:Effect-medium	-0.215	0.0110	-19.5	< 0.001	***
SampSize-8:DataPrep-post-exclusion:Effect-medium	-0.198	0.0110	-18.1	< 0.001	***
SampSize-8:DataPrep-residualized:Effect-medium	0.0676	0.0117	5.78	7.7e-09	***
SampSize-8:DataPrep-pre-exclusion:Anlys-stndrd	-0.121	0.0110	-11.0	< 0.001	***
SampSize-8:DataPrep-post-exclusion:Anlys-stndrd	-0.117	0.0110	-10.6	< 0.001	***
SampSize-8:DataPrep-residualized:Anlys-stndrd	0.0564	0.0117	4.82	1.4e-06	***
SampSize-8:Effect-medium:Anlys-stndrd	-0.0839	0.00882	-9.51	< 0.001	***
DataPrep-pre-exclusion:Effect-medium:Anlys-stndrd	-0.128	0.0110	-11.6	< 0.001	***
DataPrep-post-exclusion:Effect-medium:Anlys-stndrd	-0.132	0.0110	-12.0	< 0.001	***
DataPrep-residualized:Effect-medium:Anlys-stndrd	0.0625	0.0117	5.35	9.0e-08	***

term	estimate	std.error	statistic	p.value	
SampSize-8:DataPrep-pre-exclusion:Effect-medium:Anlys-stndrd	0.0755	0.0110	6.85	7.5e-12	***
SampSize-8:DataPrep-post-exclusion:Effect-medium:Anlys-stndrd	0.0814	0.0110	7.42	1.2e-13	***
SampSize-8:DataPrep-residualized:Effect-medium:Anlys-stndrd	-0.0400	0.0117	-3.42	0.00062	***

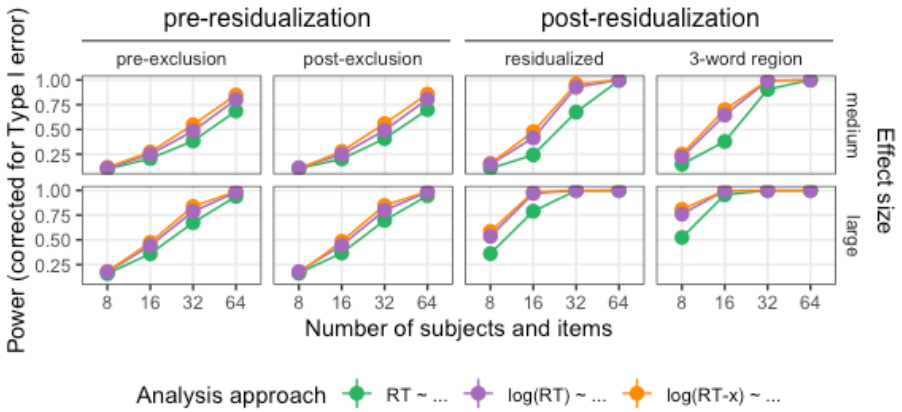

Figure 33: Power (corrected for Type I error rates) for both analysis approaches on the F13 data. Effect sizes have been collapsed into two categories: 'medium' (15 and 56 ms) and 'large' (35 and 80 ms).

NSC. The log-transformed analysis approach had higher Type I-corrected power than the untransformed approach for 31 of the 31 comparisons that were neither at ceiling nor at floor (of which there were 1). The log-transformed analysis approach had, on average, 1.95 log-odds higher Type I-corrected power (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of, e.g., 95.5% vs. 64.5% power.

The nested model comparison found that analysis approach had a significant effect on Type I-corrected power with the NSC simulations (Table 18).

Table 18: Results of nested model comparison of Type I-corrected power in NSC BATAs in Study 2. Comparing a full interaction model (i.e., sample size, data preparation, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	9	132	5.67e-24	***
2	15	9,963	0.00e+00	***
3	7	23,820	0.00e+00	***
4	1	77,373	0.00e+00	***

After removing the 32x32 and 64x64 sample size simulations to avoid singularities, the full interaction model found that the log-transformed analysis had significantly better power than the untransformed approach ($p<0.001$). Every interaction with analysis approach was significant, other than the three- and four-way interactions for the null effect, residualized data preparation condition ($ps<0.0001$).

Table 19: The summary of the full interaction model for the NSC Type I-corrected power in Study 2.

term	estimate	std.error	statistic	p.value	
(Intercept)	-0.368	0.00486	-75.8	< 0.001	***
SampSize-8	-0.640	0.00486	-132.	< 0.001	***
DataPrep-pre-exclusion	0.682	0.00743	91.7	< 0.001	***
DataPrep-post-exclusion	1.27	0.00896	142.	< 0.001	***
DataPrep-residualized	-1.13	0.00876	-129.	< 0.001	***
Effect-medium	-0.609	0.00486	-125.	< 0.001	***

term	estimate	std.error	statistic	p.value	
Anlys-stndrd	-0.372	0.00486	-76.6	< 0.001	***
SampSize-8:DataPrep-pre-exclusion	-0.0147	0.00743	-1.98	0.04740	*
SampSize-8:DataPrep-post-exclusion	-0.323	0.00896	-36.1	< 0.001	***
SampSize-8:DataPrep-residualized	0.189	0.00876	21.6	< 0.001	***
SampSize-8:Effect-medium	0.108	0.00486	22.3	< 0.001	***
DataPrep-pre-exclusion:Effect-medium	0.303	0.00743	40.7	< 0.001	***
DataPrep-post-exclusion:Effect-medium	0.0303	0.00896	3.38	0.00072	***
DataPrep-residualized:Effect-medium	-0.106	0.00876	-12.1	< 0.001	***
SampSize-8:Anlys-stndrd	0.155	0.00486	32.0	< 0.001	***
DataPrep-pre-exclusion:Anlys-stndrd	-0.0612	0.00743	-8.24	< 0.001	***
DataPrep-post-exclusion:Anlys-stndrd	-0.166	0.00896	-18.6	< 0.001	***
DataPrep-residualized:Anlys-stndrd	0.143	0.00876	16.3	< 0.001	***
Effect-medium:Anlys-stndrd	0.0629	0.00486	12.9	< 0.001	***
SampSize-8:DataPrep-pre-exclusion:Effect-medium	-0.114	0.00743	-15.3	< 0.001	***
SampSize-8:DataPrep-post-exclusion:Effect-medium	0.0276	0.00896	3.08	0.00210	**
SampSize-8:DataPrep-residualized:Effect-medium	0.0295	0.00876	3.36	0.00077	***
SampSize-8:DataPrep-pre-exclusion:Anlys-stndrd	0.0719	0.00743	9.68	< 0.001	***
SampSize-8:DataPrep-post-exclusion:Anlys-stndrd	0.0918	0.00896	10.3	< 0.001	***
SampSize-8:DataPrep-residualized:Anlys-stndrd	-0.0695	0.00876	-7.93	2.2e-15	***
SampSize-8:Effect-medium:Anlys-stndrd	-0.0302	0.00486	-6.22	5.1e-10	***
DataPrep-pre-exclusion:Effect-medium:Anlys-stndrd	-0.0687	0.00743	-9.24	< 0.001	***
DataPrep-post-exclusion:Effect-medium:Anlys-stndrd	0.0296	0.00896	3.30	0.00097	***
DataPrep-residualized:Effect-medium:Anlys-stndrd	0.0118	0.00876	1.35	0.17817	
SampSize-8:DataPrep-pre-exclusion:Effect-medium:Anlys-stndrd	0.0295	0.00743	3.97	7.1e-05	***
SampSize-8:DataPrep-post-exclusion:Effect-medium:Anlys-stndrd	-0.0229	0.00896	-2.56	0.01041	*
SampSize-8:DataPrep-residualized:Effect-medium:Anlys-stndrd	-0.0017	0.00876	-0.190	0.84924	

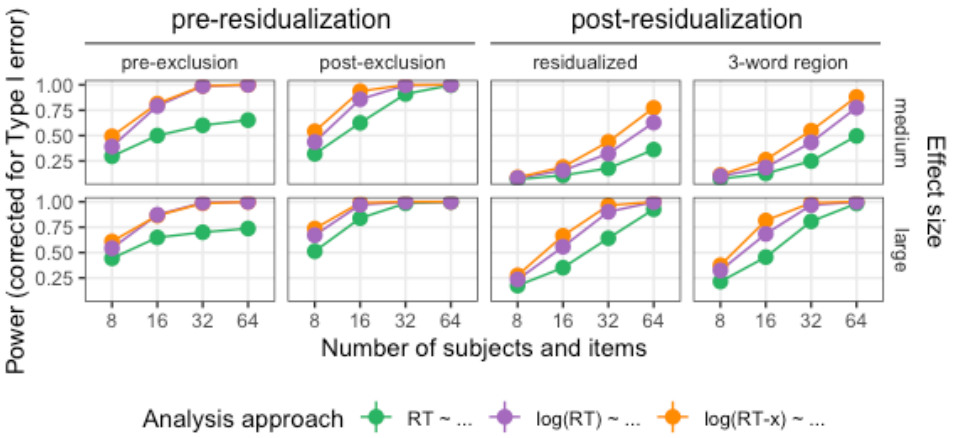

Figure 34: Power (corrected for Type I error rates) for both analysis approaches on the NSC data. Effect sizes have been collapsed into two categories: 'medium' (15 and 56 ms) and 'large' (35 and 80 ms).

SI-4 Auxiliary Study 2a—Number of sampled items for HS18

Auxiliary Study 2a investigates whether designs in which the number of subjects and items were not identical replicate the results of Study 2. We find that the number of items has little effect on convergence rates, rates of singular fit, Type I error rates, or corrected power. The likely reason for this is that by-subject variability (measured as SDs of by-subject random intercepts) was approximately four times the size of by-item variability (SDs of by-item random intercepts) in the HS18 source data.

SI-4.1 Data and simulation conditions

We limit the simulations for Study 2a to the post-exclusion condition of the HS18 source data. Additionally, we only simulate medium effect sizes for the power analysis.

SI-4.2 Results

All of these additional simulations replicate the results of Study 2, including the power advantage of the log-RT analysis (Figure 35).

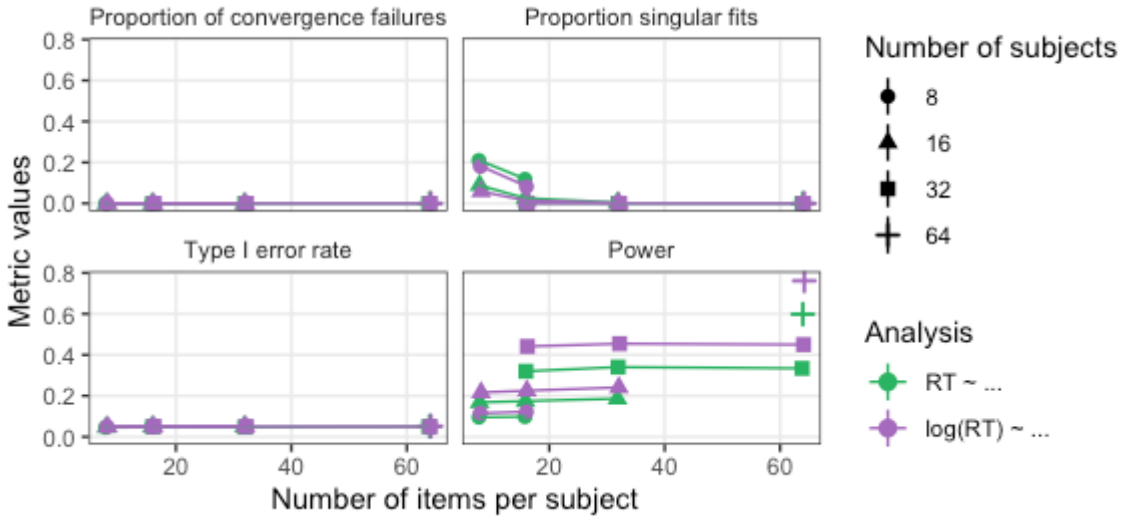

Figure 35. The convergence rates, singular fit rates, Type I error rate, and corrected power for the post-exclusion HS18 data, across BATAs created using five different subject-to-item ratios (a ratio of 1:1 is used in the main text). 95% CIs of binomial score-test are plotted but too small to be visible.

SI-5 Auxiliary Study 2b—Filler-to-item ratio for residualization

Study 2 employed a filler-to-item ratio of 15:1 for residualization (i.e., when creating BATAs, we sample 15 times as many filler RTs as critical item RTs). In order to assess whether our results depend on this choice, we conduct additional simulations with other filler-to-item ratios. We find that these ratios have little effect on convergence rates, rates of singular fit, Type I error rates, or corrected power.

SI-5.1 Data and simulation conditions

We limit ourselves to the residualized condition of the HS18 source data (filler-to-item ratios cannot affect data preparation steps prior to residualization). Additionally, we only simulate medium effect sizes crossed with the smallest and largest sample size (8 or 64 subject and items).

SI-5.2 Results

All of these additional simulations replicate the results of Study 2, including the power advantage of the log-RT analysis (Figure 36).

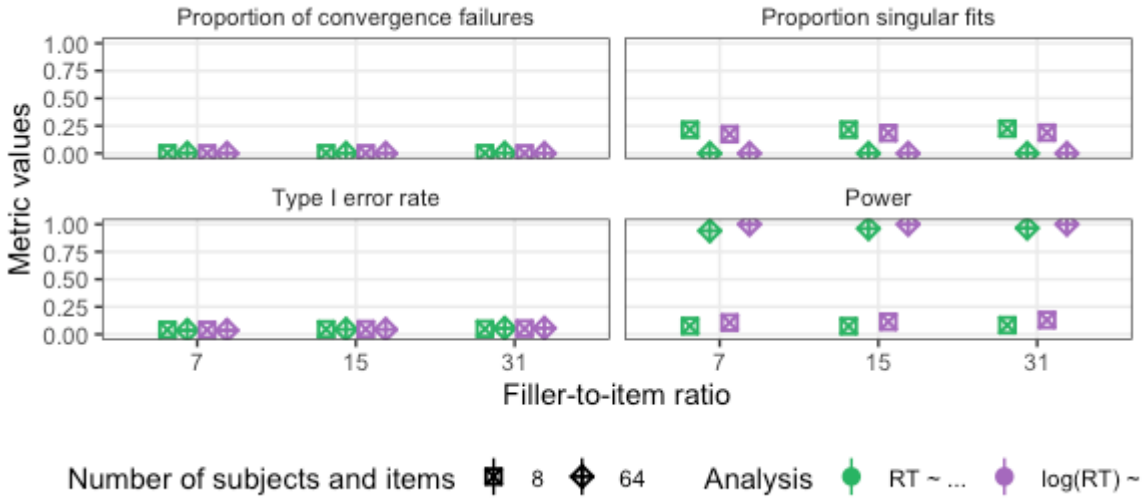

Figure 36: The convergence rates, singular fit rates, Type I error rates, and corrected power for the residualized HS18 data, across BATAs created using three different filler-to-item ratios (a ratio of 15:1 is used in the main text). 95% CIs of binomial score-test are plotted but too small to be visible.

SI-6 Auxiliary Study 2c—Number of sampled stories for NSC

Study 2 presented in the main text sampled 4 stories per BATA from the NSC. In order to assess whether our results depend on this choice, we conduct additional simulations that varied the number of sampled stories: two, four, and eight. The Type I error rates and corrected power for all three values are shown below. BATAs with eight stories and eight items were not simulated: the sampling structure for NSC BATAs manipulates the main effect between items within each story, which is not possible with a single item per story. We find that the number of stories sampled per BATA has little effect on convergence rates, rates of singular fit, Type I error rates, or corrected power.

SI-6.1 Data and simulation conditions

We limit the simulations for Study 2c to the post-exclusion condition of the NSC source data. Additionally, we only simulate medium effect sizes for the power analysis.

SI-6.2 Results

All of these additional simulations replicate the results of Study 2, including the power advantage of the log-RT analysis (Figure 37).

Figure 37. The convergence rates, singular fit rates, Type I error rate, and corrected power for the post-exclusion NSC data, across BATAs created using three different numbers of sampled stories per BATA. 95% CIs of binomial score-test are plotted, but too small to be visible.

SI-7 Auxiliary Study 2d—Adding effect in raw or log-RTs and to different mean-RTs

Study 2d introduces and crosses two new simulation conditions that further assess the robustness of the findings we obtained in Study 2. First, we assess whether the result of Study 2 replicate for different subsets of data that are selected to differ in their mean RTs. This sheds further light on the generalizability of the results of Study 2.

It also addresses a practical question for the design and analysis of SPR experiments. When participants are first faced with the unfamiliar SPR task, they exhibit slow RTs; with increasing familiarity, RTs decrease and approach a lower bound (recall Figure 20 in SI-1). This changes in the overall reading speed are expected to correlate with changes in RT variability: in SI-2, we showed that the larger mean RTs generally imply larger standard deviations around those means. This raises the question of whether the advantage of the log-transformed is limited to, for example, the early parts of the experiment (when variability is largest) or the later parts of the experiment (when participants have more or less converged against the lower bound, which is expected to reduce variability in their RTs). Alternatively, it is possible that the advantage of the log-transformed approach holds across the entire experiment. To better understand how the power advantage of the log-transformed approach in Study 2 depends on trial position and mean RTs, Study 2d compares the untransformed and log-transformed analysis approaches across BATAs generated from subsets of source data that vary in their mean RTs from fast to slow.

Second, we address the consequences of the assumption mentioned in the previous paragraph: that effects are linear in raw RTs. This is the assumption implicitly made in any analyses over untransformed RTs, i.e., the assumption made in *most* RT analyses to this day. Study 2 generated BATAs to match this assumption (as did auxiliary Studies 2a-c). Study 2d crosses the mean-RT simulation conditions with the scale in which effects are assumed to be linear: raw or log-RTs (see also Study 4 on the main text).[22]

[22] Study 2d is more directly comparable to Study 2: whereas Study 4 employs a 2x2 design, Study 2d employs the same between-subject design as Studies 2.

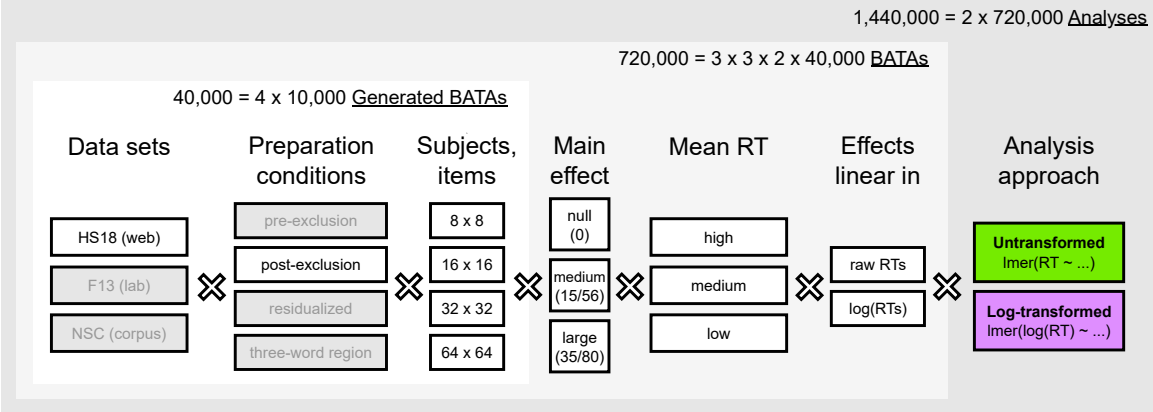

Figure 38. The design of Auxiliary Study 2d. Simulation conditions present in Study 2, but not included in Study 2d, are grayed out.

SI-7.1 Data and simulation conditions

We limit ourselves to the post-exclusion condition of the HS18 source data. The two new simulation conditions crossed in Study 2d are described next.

Mean-RTs. To assess whether the results of Study 2 hold across mean-RTs, we sample RTs in three different conditions: low, medium, and high mean RTs. This is achieved by sampling RTs from one of three different parts of the HS18 experiment, each corresponding to five adjacent filler trials at the beginning (high RT mean), middle (medium RTs mean), and end of the experiment (low RT mean). Table 20 summarizes the mean RTs and other descriptive statistics of the three conditions. Figure 39 visualizes the RT distribution for each mean-RT condition.

Table 20: Descriptive statistics of the three mean-RT conditions from the HS18 data that are used for Study 2d. The "high" mean-RT data come from the early trials in the original experiment by Harrington Stack et al. (2018), the "medium" mean-RT data come from trials near the middle, and the "low" mean-RT data come from trials near the end.

Mean RT bin	mean	SD	skewness	excess kurtosis
low	299	136	3.05	18.5
medium	338	165	2.71	13.4
high	395	204	2.41	9.9

As is evident from Table 20, RTs decrease over the course of the experiment (see also Figure 20). With decreases in mean RTs, standard deviations of those RTs also decreased. Interestingly, this is paired with an *in*crease in skewness and kurtosis, presumably as RTs begin to be "pushed" against the lower bound (Figure 39).

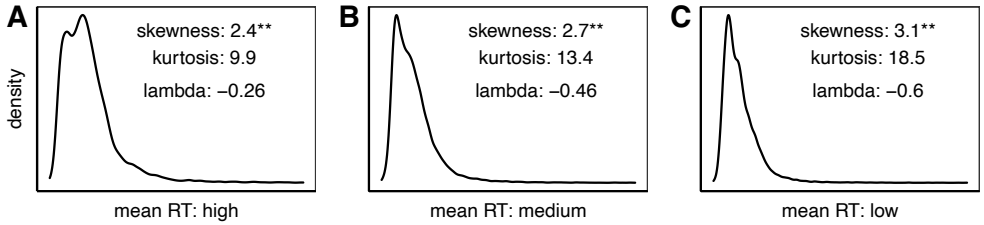

Figure 39. Marginal density of word-by-word reading times using the post-exclusion HS18 data at the three mean RTs of Study 4 and Study 2d. (See Figure 1 for details). As the mean decreases, skewness and kurtosis increase. Study 4 uses the high and low mean-RTs, while Study 2d uses all three.

Effect scale. The second new simulations condition of Study 2d—whether effects are generated to be linear in raw vs. log-RTs—is described in Study 4. For reasons discussed in Study 4, we note that a direct comparison of statistical power for effects generated to be linear in raw vs. log-RTs is not meaningful.[23] Rather, we are interested in whether the same qualitative pattern of results is observed, regardless of the scale in which the simulated effects are linear.

SI-7.2 Convergence failures

To compare the rate of convergence failure, we used a full interaction model with mean RT, sample size, effect size (null, medium, or large), and analysis approach.

Effects generated to be linear in raw RTs. Model comparison did not find any significant effect of analysis approach on rate of convergence failures (Table 21).

Table 21: Results of nested model comparison of convergence failures in Study 2d when effects are linear in raw RTs. Comparing a full interaction model (i.e., sample size, mean RT, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*
1	12	1.280	1.000
2	16	2.853	1.000
3	7	1.886	0.966
4	1	0.527	0.468

Effects generated to be linear in log-RTs. Model comparison did not find any significant effect of analysis approach on rate of convergence failures (Table 22).

Table 22: Results of nested model comparison of convergence failures in Study 2d when effects are linear in log-RTs. Comparing a full interaction model (i.e., sample size, mean RT, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*
1	12	2.226	0.999
2	16	1.661	1.000
3	7	0.929	0.996
4	1	0.712	0.399

[23] The size of the effect added to log-RTs was selected to be equivalent (in raw RTs) to the effect added to raw RTs when added to the medium mean-RT condition. However, even with this constraint, adding a constant effect to log-RTs has different consequences in raw RTs, depending on the log-RT value the effect is added to. Even for the medium mean-RT condition, there is thus no guarantee that the marginal effects (averaging across all sampled tokens in a BATA) were identical across the two scales. The same concern does not apply to Type I error results, since they add 0 effects.

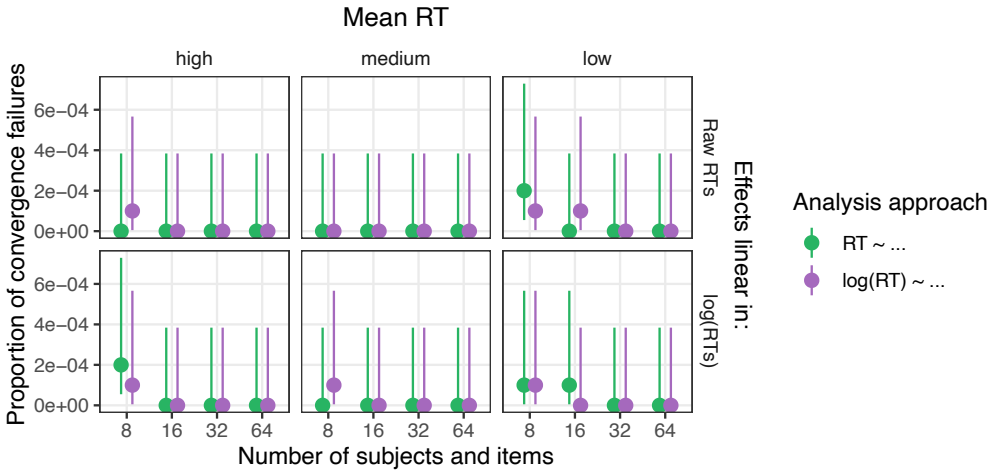

Figure 40: Convergence rates for the mean RT data in Study 2d. The CIs are 95% binomial score-test-based confidence intervals.

SI-7.3 Singular fits

Effects generated to be linear in raw RTs. Model comparison found that analysis approach had a significant effect on rates of singular fit (Table 23).

Table 23: Results of nested model comparison of singular fit rates in Study 2d when effects are linear in raw RTs. Comparing a full interaction model (i.e., sample size, mean RT, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	12	0.119	1.00e+00	
2	16	100.570	2.71e-14	***
3	7	499.461	1.05e-103	***
4	1	318.318	3.37e-71	***

After removing the 64x64 sample size simulations to avoid singularities, the full interaction model found that the log-transformed analysis had significantly fewer singular fits than the untransformed approach ($p<0.001$). It also had a number of significant interactions with mean-RT and sample size, including some higher-order interactions.

Effects generated to be linear in log-RTs. Model comparison found that analysis approach had a significant effect on rates of singular fit (Table 24).

Table 24: Results of nested model comparison of singular fit rates in Study 2d when effects are linear in log-RTs. Comparing a full interaction model (i.e., sample size, mean RT, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	12	0.279	1.00e+00	
2	16	104.913	4.12e-15	***
3	7	595.964	1.80e-124	***
4	1	358.036	7.54e-80	***

After removing the 64x64 sample size simulations to avoid singularities, the full interaction model found that the log-transformed analysis had significantly fewer singular fits than the untransformed approach ($p<0.001$). It also had a number of significant interactions with mean-RT and sample size, including some higher-order interactions.

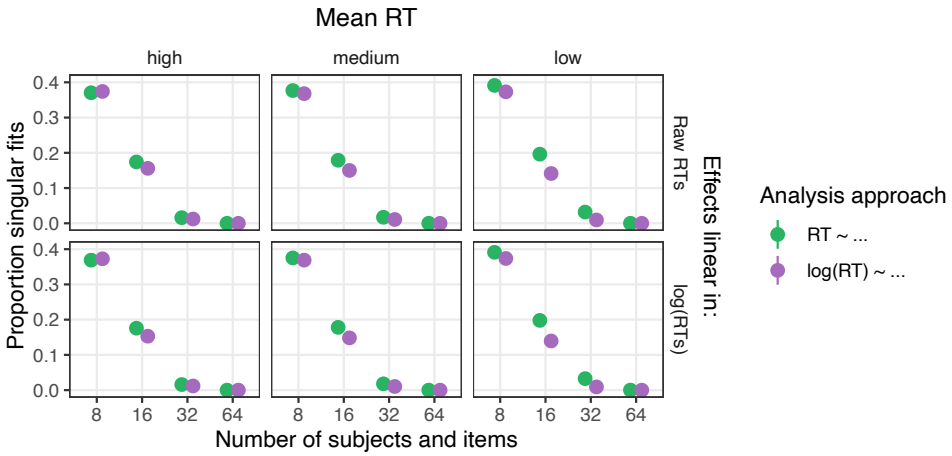

Figure 41: Proportion of singular fits for the mean RT data in Study 2d. 95% CIs of binomial score-test are plotted, but too small to be visible.

SI-7.4 Type I errors

Replicating the HS18 post-exclusion condition of Study 2, the Type I error rates for the different HS18 mean-RTs were close to the targeted value of 0.05 (Figure 42). Also replicating Study 2, the untransformed analysis approach had more simulations with Type I error rates significantly *below* the expected 0.05 rate, which suggests that the untransformed approach can be conservative (significant anti-conservativity was not observed in any of the simulation conditions). The log-transformed analysis approach had Type I error rates closer to the expected value for 16 of the 24 simulations.

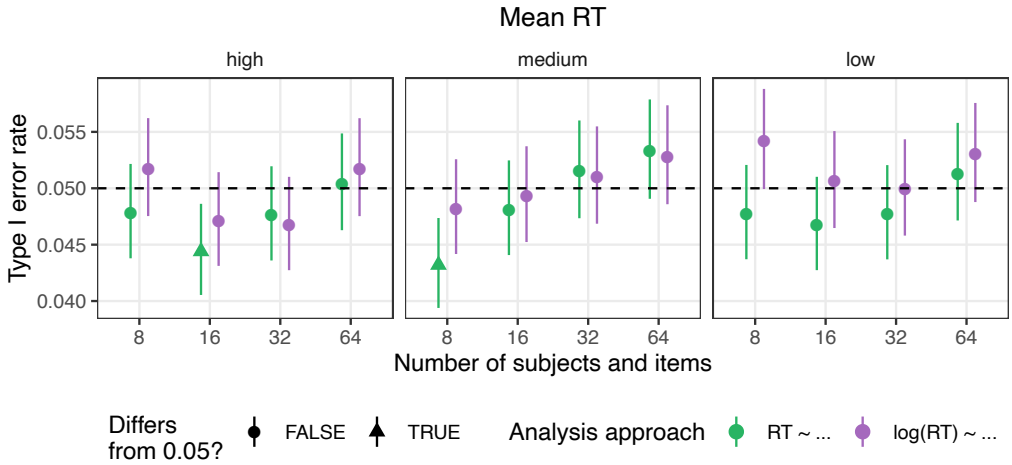

Figure 42. The Type I error rates in Study 2d. Type I error rates were generally at the expected rate of 0.05, except for two instances of the untransformed approach that resulted in conservative Type I error rates.

Since null effects are linear in both raw RTs and log-RTs, only one analysis is needed. Model comparison found that analysis approach had a significant effect on Type I error rates (Table 25).

Table 25: Results of nested model comparison of Type I error rates in Study 2d. Comparing a full interaction model (i.e., sample size, mean RT, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	6	0.517	0.9976	
2	5	5.852	0.3209	
3	1	6.221	0.0126	*

Since Type I error rates never hit ceiling or floor, no data needed to be removed to prevent singularities in the logistic regression. The full interaction model found that the log-transformed analysis had significantly higher Type I error rates than the untransformed approach (p=0.012). The effect of analysis approach had one interaction with sample size, with the difference between approaches being greater for the 8x8 BATAs.

Table 26: The summary of the full interaction model for the Type I data in Study 2d.

term	estimate	std.error	statistic	p.value	
(Intercept)	-2.96	0.00943	-314.	<2e-16	***
SampSize-8	-0.0140	0.0164	-0.853	0.3938	
SampSize-16	-0.0363	0.0165	-2.20	0.0279	*
SampSize-32	-0.00650	0.0164	-0.397	0.6913	
MeanRT-high	-0.0209	0.0134	-1.56	0.1189	
MeanRT-medium	0.00520	0.0133	0.390	0.6966	
Anlys-stndrd	-0.0238	0.00943	-2.53	0.0115	*
SampSize-8:MeanRT-high	0.0429	0.0232	1.85	0.0640	
SampSize-16:MeanRT-high	-0.0225	0.0235	-0.954	0.3402	
SampSize-32:MeanRT-high	-0.0203	0.0233	-0.872	0.3833	
SampSize-8:MeanRT-medium	-0.0731	0.0234	-3.12	0.0018	**
SampSize-16:MeanRT-medium	0.0174	0.0233	0.747	0.4548	
SampSize-32:MeanRT-medium	0.0414	0.0230	1.80	0.0719	
SampSize-8:Anlys-stndrd	-0.0315	0.0164	-1.92	0.0553	
SampSize-16:Anlys-stndrd	-0.00483	0.0165	-0.293	0.7698	
SampSize-32:Anlys-stndrd	0.0210	0.0164	1.28	0.1994	
MeanRT-high:Anlys-stndrd	0.00493	0.0134	0.368	0.7130	
MeanRT-medium:Anlys-stndrd	0.00878	0.0133	0.659	0.5101	
SampSize-8:MeanRT-high:Anlys-stndrd	0.00903	0.0232	0.390	0.6966	
SampSize-16:MeanRT-high:Anlys-stndrd	-0.00715	0.0235	-0.304	0.7615	
SampSize-32:MeanRT-high:Anlys-stndrd	0.00792	0.0233	0.339	0.7343	
SampSize-8:MeanRT-medium:Anlys-stndrd	-0.0109	0.0234	-0.464	0.6427	
SampSize-16:MeanRT-medium:Anlys-stndrd	0.00688	0.0233	0.296	0.7674	
SampSize-32:MeanRT-medium:Anlys-stndrd	-0.000752	0.0230	-0.0327	0.9739	

SI-7.5 Power (uncorrected)

Effects generated to be linear in raw RTs. For the mean-RT simulations with effects generated to be linear in raw RTs, the log-transformed analysis approach had higher power than the standard approach for 24 of the 24 comparisons that were not both at ceiling or floor. The log-transformed analysis approach had, on average, 0.492 log-odds higher power (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of 75.8% vs. 83.6%.

Effects generated to be linear in log-RTs. For the mean-RT simulations with effects generated to be linear in log-RTs, the log-transformed analysis approach had higher power than the standard approach for 24 of the 24 comparisons that were not both at ceiling or floor. The log-transformed analysis approach had, on average, 0.138 log-odds higher power (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of, e.g., 81.1% vs. 78.9% power. While this power advantage is much less

pronounced than for effects that were generated to be linear in raw RTs, we remind readers that those advantages are not directly comparable, as the effect sizes we added to raw and log-RTs are not directly comparable (see footnote 23).

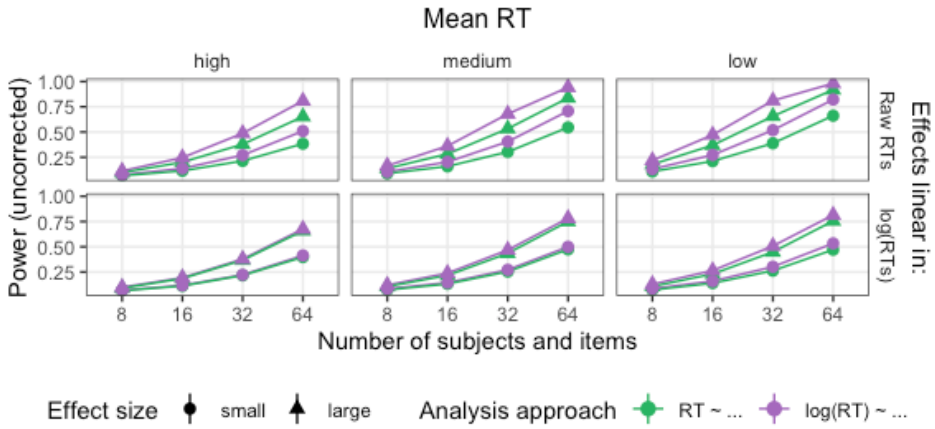

Figure 43: Power (uncorrected) for both analysis approaches in Study 2d. 95% CIs of binomial score-test are plotted, but too small to be visible.

SI-7.6 Power (corrected for Type I error rate)

Unsurprisingly, and replicating Study 2, power increased along with sample and effect sizes (Figure 44). Power also increased as mean RTs decreased, with the low mean-RT (coming from the end of the original experiment) having the most power, regardless of analysis approach. This is expected, given that SDs decreased along with mean RTs (see Table 20).

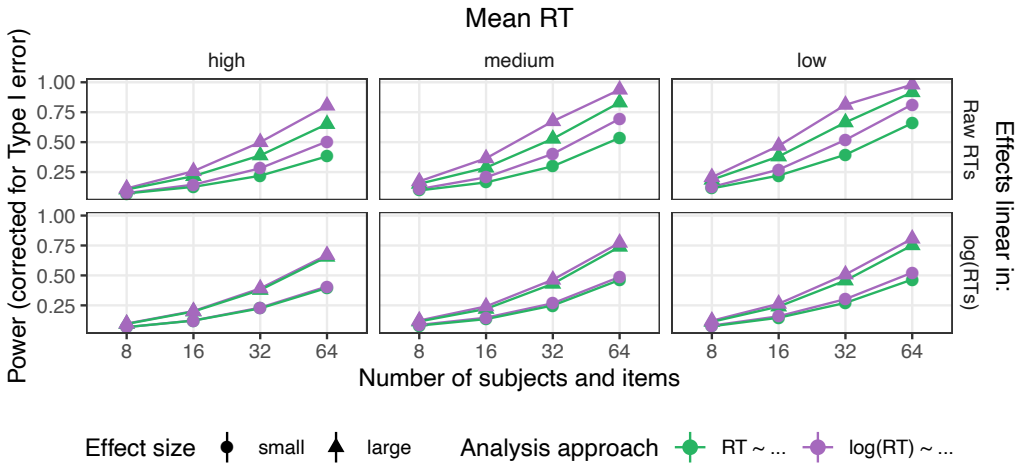

Figure 44. Power (corrected for Type I error rates) Study 2d. 95% CIs of binomial score-test are plotted, but too small to be visible.

Figure 45 shows the power advantage of the log-transformed analysis, compared to the untransformed approach, depending on the simulation conditions. Two results are immediately obvious. First, as the sample size increased, so did the power advantage of log-transformed analysis, replicating Study 2 across a range of different mean RTs.

Second, as the mean-RT decreased, the power advantage of the log-transformed analysis *in*creased. This is perhaps surprising, as one might initially consider the presence of particularly *slow* trials (e.g., outliers due to lapses in attention) the reason why the log-transformed analysis approach outperforms the

untransformed approach in terms of power. Study 2d suggests that this is not the case. Rather, it seems to be the presence of a soft lower bound that drives the power advantage of the log-transformed approach. This would also suggest that the disadvantage of the untransformed approach increases for experiments with RTs closer to the soft floor for reading times—for example, particularly long experiments (as is the case for the HS18 data set).

There is, however, an alternative explanation for the increasing power advantage for faster mean-RTs: because of the positive correlation between the mean and SD of RTs, the smaller the mean-RT, the less variability there is around that mean. This means that *effective* effect size increases for faster RTs. In both Study 2 and 2d, we find that the power advantage of the log-transformed approach increases for larger effect sizes. It is thus possible that the increasing power advantage of the log-transformed approach for faster mean-RTs in Figure 45 is entirely driven by its increasing power advantage for larger (effective) effect sizes.

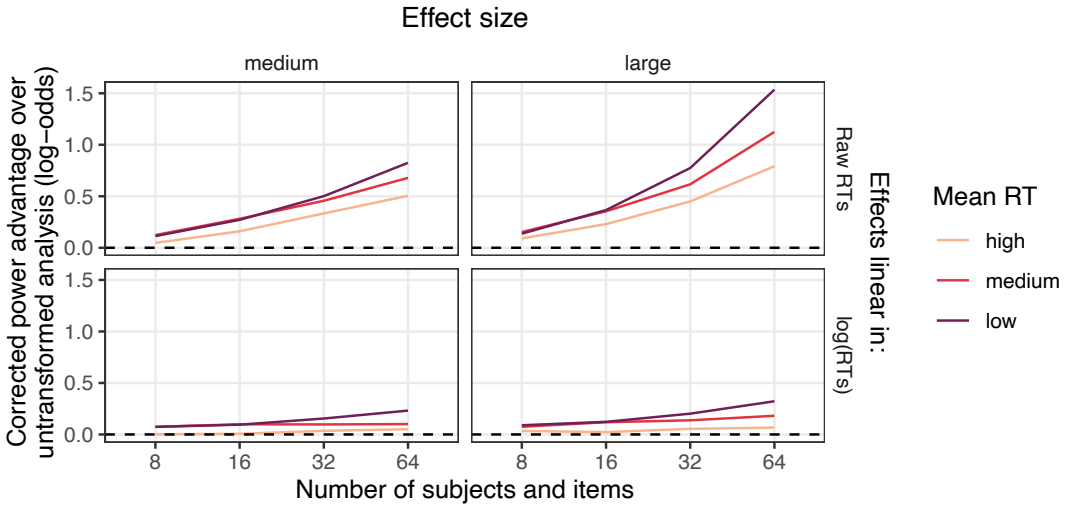

Figure 45. Relative power advantage of the log-transformed approach in Study 2d. Type I error-corrected relative power advantage of the log-transformed analysis approach compared to the untransformed analysis approach (as a difference in log-odds) for all three mean-RTs and each effect size. Simulations for which power was at ceiling in either analysis approach are excluded (because they would result in infinite log-odds).

Effects generated to be linear in raw RTs. Model comparison found that analysis approach had a significant effect on Type I-corrected power (Table 27).

Table 27: Results of nested model comparison of Type I-corrected power in Study 2d when effects are linear in raw RTs. Comparing a full interaction model (i.e., sample size, mean RT, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	6	14.9	2.09e-02	*
2	11	124.8	2.00e-21	***
3	6	1,245.1	8.46e-266	***
4	1	4,180.8	0.00e+00	***

The full interaction model found that the log-transformed analysis had significantly better power than the untransformed approach ($p<0.001$).

Table 28: The summary of the full interaction model for the Type I-corrected power in Study 2d when effects are linear in raw RTs.

term	estimate	std.error	statistic	p.value	
(Intercept)	-0.481	0.00402	-119.	< 0.001	***
SampSize-8	-1.50	0.00758	-198.	< 0.001	***
SampSize-16	-0.633	0.00636	-99.4	< 0.001	***
SampSize-32	0.375	0.00596	62.9	< 0.001	***
MeanRT-high	-0.546	0.00560	-97.6	< 0.001	***
MeanRT-medium	0.0299	0.00552	5.41	6.2e-08	***
Effect-medium	-0.500	0.00402	-124.	< 0.001	***
Anlys-stndrd	-0.227	0.00402	-56.5	< 0.001	***
SampSize-8:MeanRT-high	0.201	0.0112	18.0	< 0.001	***
SampSize-16:MeanRT-high	0.146	0.00923	15.8	< 0.001	***
SampSize-32:MeanRT-high	-0.0114	0.00839	-1.36	0.17438	
SampSize-8:MeanRT-medium	0.0405	0.0105	3.86	0.00011	***
SampSize-16:MeanRT-medium	-0.0267	0.00885	-3.02	0.00255	**
SampSize-32:MeanRT-medium	-0.0258	0.00825	-3.13	0.00174	**
SampSize-8:Effect-medium	0.241	0.00758	31.8	< 0.001	***
SampSize-16:Effect-medium	0.124	0.00636	19.5	< 0.001	***
SampSize-32:Effect-medium	-0.0283	0.00596	-4.74	2.1e-06	***
MeanRT-high:Effect-medium	0.0945	0.00560	16.9	< 0.001	***
MeanRT-medium:Effect-medium	-0.00047	0.00552	-0.086	0.93180	
SampSize-8:Anlys-stndrd	0.172	0.00758	22.7	< 0.001	***
SampSize-16:Anlys-stndrd	0.0889	0.00636	14.0	< 0.001	***
SampSize-32:Anlys-stndrd	-0.0336	0.00596	-5.64	1.7e-08	***
MeanRT-high:Anlys-stndrd	0.0644	0.00560	11.5	< 0.001	***
MeanRT-medium:Anlys-stndrd	-0.00932	0.00552	-1.69	0.09124	
Effect-medium:Anlys-stndrd	0.0484	0.00402	12.0	< 0.001	***
SampSize-8:MeanRT-high:Effect-medium	-0.0533	0.0112	-4.77	1.8e-06	***
SampSize-16:MeanRT-high:Effect-medium	-0.0590	0.00923	-6.39	1.6e-10	***
SampSize-32:MeanRT-high:Effect-medium	-0.00419	0.00839	-0.500	0.61733	
SampSize-8:MeanRT-medium:Effect-medium	-0.00501	0.0105	-0.479	0.63222	
SampSize-16:MeanRT-medium:Effect-medium	-0.00170	0.00885	-0.192	0.84745	
SampSize-32:MeanRT-medium:Effect-medium	0.00582	0.00825	0.706	0.48033	
SampSize-8:MeanRT-high:Anlys-stndrd	-0.0436	0.0112	-3.90	9.5e-05	***
SampSize-16:MeanRT-high:Anlys-stndrd	-0.0232	0.00923	-2.52	0.01184	*
SampSize-32:MeanRT-high:Anlys-stndrd	0.000648	0.00839	0.0772	0.93850	
SampSize-8:MeanRT-medium:Anlys-stndrd	-0.00433	0.0105	-0.414	0.67913	
SampSize-16:MeanRT-medium:Anlys-stndrd	-0.0110	0.00885	-1.25	0.21196	
SampSize-32:MeanRT-medium:Anlys-stndrd	0.00221	0.00825	0.268	0.78902	
SampSize-8:Effect-medium:Anlys-stndrd	-0.0403	0.00758	-5.32	1.0e-07	***
SampSize-16:Effect-medium:Anlys-stndrd	-0.0286	0.00636	-4.50	6.8e-06	***
SampSize-32:Effect-medium:Anlys-stndrd	-0.00259	0.00596	-0.435	0.66335	
MeanRT-high:Effect-medium:Anlys-stndrd	-0.0161	0.00560	-2.87	0.00409	**
MeanRT-medium:Effect-medium:Anlys-stndrd	-0.00415	0.00552	-0.752	0.45218	
SampSize-8:MeanRT-high:Effect-medium:Anlys-stndrd	0.0189	0.0112	1.69	0.09088	
SampSize-16:MeanRT-high:Effect-medium:Anlys-stndrd	0.0137	0.00923	1.49	0.13615	
SampSize-32:MeanRT-high:Effect-medium:Anlys-stndrd	-0.00054	0.00839	-0.064	0.94894	
SampSize-8:MeanRT-medium:Effect-medium:Anlys-stndrd	0.00353	0.0105	0.337	0.73643	
SampSize-16:MeanRT-medium:Effect-medium:Anlys-stndrd	0.00251	0.00885	0.284	0.77638	
SampSize-32:MeanRT-medium:Effect-medium:Anlys-stndrd	-0.00171	0.00825	-0.207	0.83588	

Effects generated to be linear in log-RTs. Model comparison found that analysis approach had a significant effect on Type I-corrected power (Table 29).

Table 29: Results of nested model comparison of Type I-corrected power in Study 2d when effects are linear in log-RTs. Comparing a full interaction model (i.e., sample size, mean RT, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	6	1.09	9.82e-01	
2	11	15.73	1.52e-01	
3	6	93.47	5.75e-18	***
4	1	258.92	2.94e-58	***

The full interaction model found that the log-transformed analysis had significantly better power than the untransformed approach ($p<0.001$).

Table 30: The summary of the full interaction model for the Type I-corrected power in Study 2d when effects are linear in log-RTs.

term	estimate	std.error	statistic	p.value	
(Intercept)	-1.01	0.00386	-262.	< 0.001	***
SampSize-8	-1.28	0.00812	-157.	< 0.001	***
SampSize-16	-0.516	0.00666	-77.4	< 0.001	***
SampSize-32	0.355	0.00587	60.5	< 0.001	***
MeanRT-high	-0.190	0.00556	-34.2	< 0.001	***
MeanRT-medium	0.0571	0.00541	10.5	< 0.001	***
Effect-medium	-0.377	0.00386	-97.7	< 0.001	***
Anlys-stndrd	-0.0509	0.00386	-13.2	< 0.001	***
SampSize-8:MeanRT-high	0.0519	0.0118	4.41	1.0e-05	***
SampSize-16:MeanRT-high	0.0429	0.00963	4.46	8.2e-06	***
SampSize-32:MeanRT-high	0.00227	0.00843	0.269	0.78809	
SampSize-8:MeanRT-medium	0.0370	0.0113	3.29	0.00102	**
SampSize-16:MeanRT-medium	-0.0349	0.00937	-3.73	0.00019	***
SampSize-32:MeanRT-medium	-0.0345	0.00825	-4.18	2.9e-05	***
SampSize-8:Effect-medium	0.175	0.00812	21.5	< 0.001	***
SampSize-16:Effect-medium	0.0765	0.00666	11.5	< 0.001	***
SampSize-32:Effect-medium	-0.0289	0.00587	-4.92	8.7e-07	***
MeanRT-high:Effect-medium	0.0299	0.00556	5.39	7.2e-08	***
MeanRT-medium:Effect-medium	-0.00712	0.00541	-1.32	0.18848	
SampSize-8:Anlys-stndrd	0.0220	0.00812	2.71	0.00663	**
SampSize-16:Anlys-stndrd	0.0123	0.00666	1.84	0.06591	
SampSize-32:Anlys-stndrd	-0.00582	0.00587	-0.992	0.32127	
MeanRT-high:Anlys-stndrd	0.0340	0.00556	6.11	9.9e-10	***
MeanRT-medium:Anlys-stndrd	-0.00424	0.00541	-0.783	0.43353	
Effect-medium:Anlys-stndrd	0.00848	0.00386	2.19	0.02818	*
SampSize-8:MeanRT-high:Effect-medium	-0.0145	0.0118	-1.23	0.21830	
SampSize-16:MeanRT-high:Effect-medium	-0.0187	0.00963	-1.94	0.05185	
SampSize-32:MeanRT-high:Effect-medium	0.000735	0.00843	0.0872	0.93052	
SampSize-8:MeanRT-medium:Effect-medium	0.0112	0.0113	0.993	0.32090	
SampSize-16:MeanRT-medium:Effect-medium	0.00648	0.00937	0.692	0.48895	
SampSize-32:MeanRT-medium:Effect-medium	-0.00805	0.00825	-0.975	0.32932	
SampSize-8:MeanRT-high:Anlys-stndrd	-0.0138	0.0118	-1.17	0.24223	
SampSize-16:MeanRT-high:Anlys-stndrd	-0.00268	0.00963	-0.279	0.78046	
SampSize-32:MeanRT-high:Anlys-stndrd	0.000364	0.00843	0.0432	0.96555	
SampSize-8:MeanRT-medium:Anlys-stndrd	-0.00419	0.0113	-0.372	0.71005	
SampSize-16:MeanRT-medium:Anlys-stndrd	-0.0113	0.00937	-1.21	0.22686	
SampSize-32:MeanRT-medium:Anlys-stndrd	0.00232	0.00825	0.281	0.77869	
SampSize-8:Effect-medium:Anlys-stndrd	-0.00448	0.00812	-0.552	0.58122	
SampSize-16:Effect-medium:Anlys-stndrd	-0.00308	0.00666	-0.462	0.64411	

term	estimate	std.error	statistic	p.value
SampSize-32:Effect-medium:Anlys-stndrd	0.000442	0.00587	0.0753	0.93998
MeanRT-high:Effect-medium:Anlys-stndrd	-0.00325	0.00556	-0.585	0.55870
MeanRT-medium:Effect-medium:Anlys-stndrd	0.000393	0.00541	0.0726	0.94211
SampSize-8:MeanRT-high:Effect-medium:Anlys-stndrd	0.00722	0.0118	0.613	0.53977
SampSize-16:MeanRT-high:Effect-medium:Anlys-stndrd	0.00200	0.00963	0.208	0.83509
SampSize-32:MeanRT-high:Effect-medium:Anlys-stndrd	-0.00086	0.00843	-0.102	0.91912
SampSize-8:MeanRT-medium:Effect-medium:Anlys-stndrd	-0.00427	0.0113	-0.379	0.70491
SampSize-16:MeanRT-medium:Effect-medium:Anlys-stndrd	-0.00066	0.00937	-0.071	0.94359
SampSize-32:MeanRT-medium:Effect-medium:Anlys-stndrd	0.000802	0.00825	0.0972	0.92256

SI-7.7 Discussion

Like Study 2, Study 2d finds an unambiguous advantage for the log-transformed analysis approach compared to the untransformed analysis approach. In all simulations, log-transforming RTs increased the probability of detecting true effects, and although the untransformed analysis approach had slightly lower Type I error rates, this appears to be because it is *overly* conservative for the RT data here. Study 2d further finds that the power deficit found in Study 2 holds across different mean RTs (and, by extension, different parts of experiments).

SI-8 Auxiliary Study 2e—A hybrid data generation approach to simulate additional by-subject and-item variability in effects

Auxiliary Study 2e tests whether the results of Study 2 would change if random by-item and -subject variability *beyond that inherently in the bootstrapped natural RTs* is added to the simulated effects. As discussed in Study 2, the hierarchical bootstrap approach introduces natural between-subject and between-item variability into the BATAs. For the type of design simulated so far—between-subject (for HS18 and F13 source data) or between-item (for NSC source data)—this variability arguably simulates the most important sources of by-subject and -item variability that would be encountered in reading analyses.

Study 2e thus simulates a by-2 *within*-subject, *within*-item experiment design. If the findings of Study 2—no anti-conservative Type I errors and a power advantage of the log-transformed approach—replicate for this design, and do so regardless of whether additional by-subject and -item variability is added to the simulated effects, this would provide further evidence that the findings of Study 2 are robust and likely to generalize.

SI-8.1 Data and simulation conditions

We limit the simulations to the medium mean-RTs of the post-exclusion condition of the HS18 source data. This decision was made to facilitate comparison to Study 2d. We simulate data for two different sample sizes: BATAs with 32 subjects and 32 items or 64 subjects and 64 items.

As the within-subject and -item design of Study 2e yields much more power than the between-subject design used in Study 2 (and Studies 2a-b, d), we reduced the simulated effects to 7 ms (medium) and 10 ms (large). As in Study 2d, we add effects—both the fixed and random effects—to be linear in either raw or log-transformed RTs.

The scale that the effects were linear in, the main effect sizes, the different BATA sizes, and the analysis approaches were crossed with a new simulation condition: whether parametrically-generated by-subject and by-item slopes and intercepts were included or not. The latter follows the same data generation approach as in Study 2d. For the former, we add normally distributed random by-item and by-subject intercept and slopes to each BATA.

In the main text, we stress that the particular effect sizes we simulate are not inherently important, in the same way that the absolute power they result in is not inherently important. Rather, we aim to understand how the two analysis approaches differ in power (among other things) when compared against data that has the same underlying (absolute) effect. The same holds for the values of the standard deviations of the

random intercepts and slopes. We did, however, decide to set those standard deviations to plausible values. Specifically, we set the standard deviations of the random intercepts and slopes based on HS18 data ($SD_{Subject,\ Intercept}$ = 105ms; $SD_{Subject,\ Slope}$ = 15ms; $SD_{Item,\ Intercept}$ = 42ms; $SD_{Item,\ Slope}$ = 10ms).[24] For effects generated to be linear in log-RTs, the random effects were added in the following way (extending the approach used to add main effects in Studies 2d and 4): the standard deviation of each random effect was adjusted to be the same size in log-RTs as raw RTs when centered around the mean RT of the medium mean-RTs. These log-RT standard deviations were used to sample the random effects from a normal distribution in log-RTs.

As we did for Study 2d, we emphasize that comparison of power across effects added in raw vs. log-RTs is not meaningful.

SI-8.2 Results

Type I errors. Figure 46 shows the Type I error rates for all simulation conditions. Replicating Study 2, Type I error rates were slightly overly conservative across the board. Also replicating Study 2, Type I error rates for the log-transformed analysis approach were consistently higher and closer to the targeted Type I error rate of 0.05 than for the untransformed approach. Finally, extending Study 2, the presence of parametrically added random effects increased the Type I error rate compared to the simulation conditions in which no parametric random effects were added.

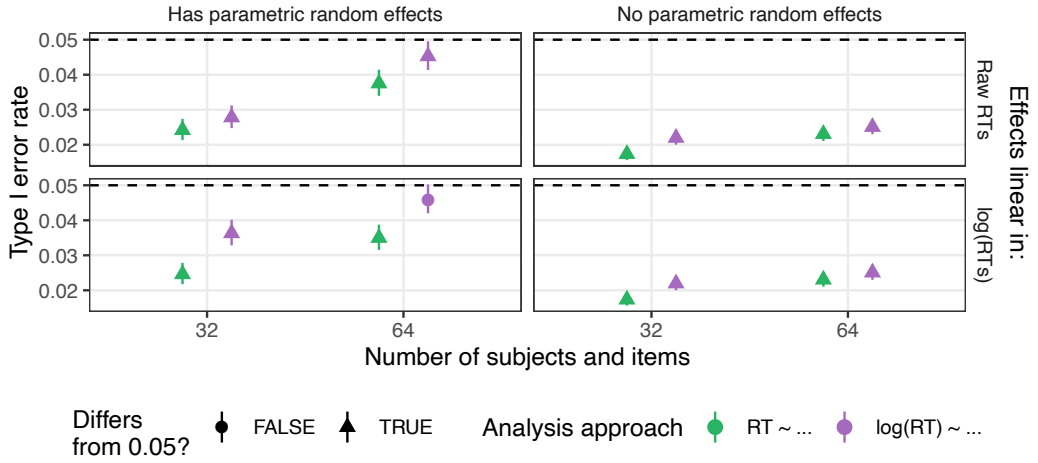

Figure 46. Type I error rates for within-subject post-exclusion HS18 BATAs, with and without added parametric by-subject and by-item random effects. Random effects were added in the scale indicated on the right. See text for details of parametric effects.

Power (corrected). Figure 47 shows the Type I corrected power for the same analyses. Replicating Study 2, in each simulation condition, the log-transformed analysis approach had higher Type I error-corrected power than the untransformed analysis approach. Extending Study 2 and Study 2d, this power advantage of the log-transformed approach held, regardless of whether random effects were included in the data generation or not. It also held for both effect sizes, and thus for random effect variability that differs relative to the absolute size of the fixed effect.

[24] These values were obtained by fitting an LMM to RTs from the critical trials of the original study—i.e., we use the post-exclusion HS18 data for the disambiguation region of relative clause stimuli for subjects in the "RC-first" group. The LMM included a main effect of ambiguity along with by-subject and by-item intercepts and slopes for the ambiguity effect.

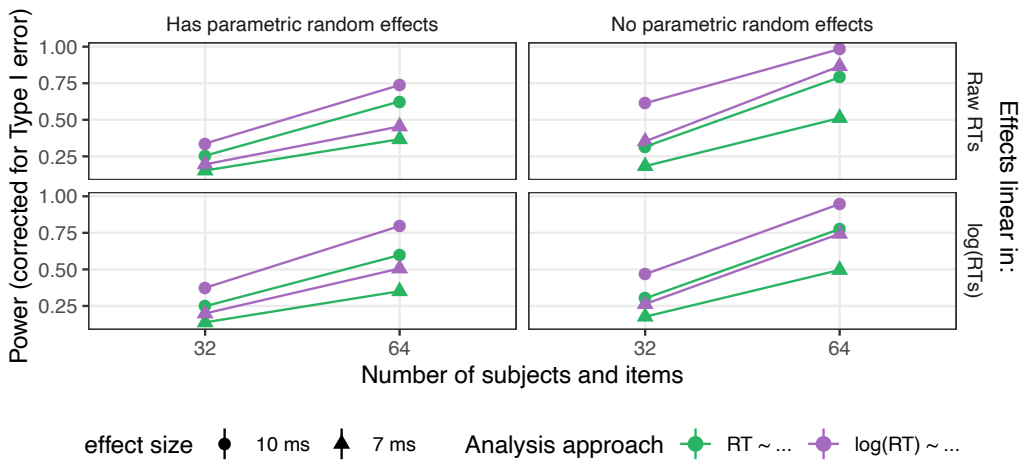

Figure 47. Type I-corrected power for the within-subject post-exclusion HS18 BATAs, with and without added parametric by-subject and by-item random effects. 95% CIs of binomial score-test are plotted, but too small to be visible.

Additional comparisons showed that this power advantage was reduced when random effects were included (see Figure 48). As already discussed for the fixed effects in Study 2d, it is, however, unclear whether this reduced advantage is simply due to the fact that power was *overall* smaller when parametric random effects were added. Figure 49 suggests as much: rather than forming outliers, the power advantage of the log-transformed approach for simulations that included random effects follows the same trend as the power advantage for simulations that did not include random effects. Future research could investigate this question further by (1) generating different fixed effect sizes for the BATAs with and without random effects that result in identical power with and without random effects for, e.g., the untransformed analysis approach, and then (2) comparing the power of the log-transformed approach over the same BATAs. Since our goal here is solely to ascertain that the qualitative power advantage of the log-transformed approach persists even when random effects are generated, we leave questions about the relative size of this advantage to future work.

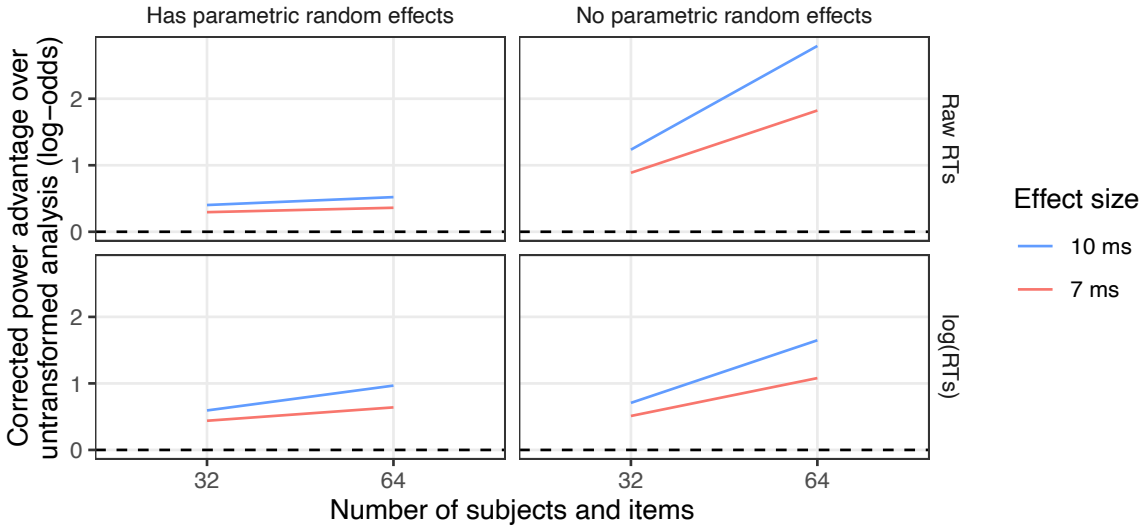

Figure 48. Relative power advantage of the log-transformed approach in Study 2e. Type I error-corrected relative power advantage of the log-transformed analysis approach compared to the untransformed

analysis approach (as a difference in log-odds) when parametric random effects were added or not and each for effect size.

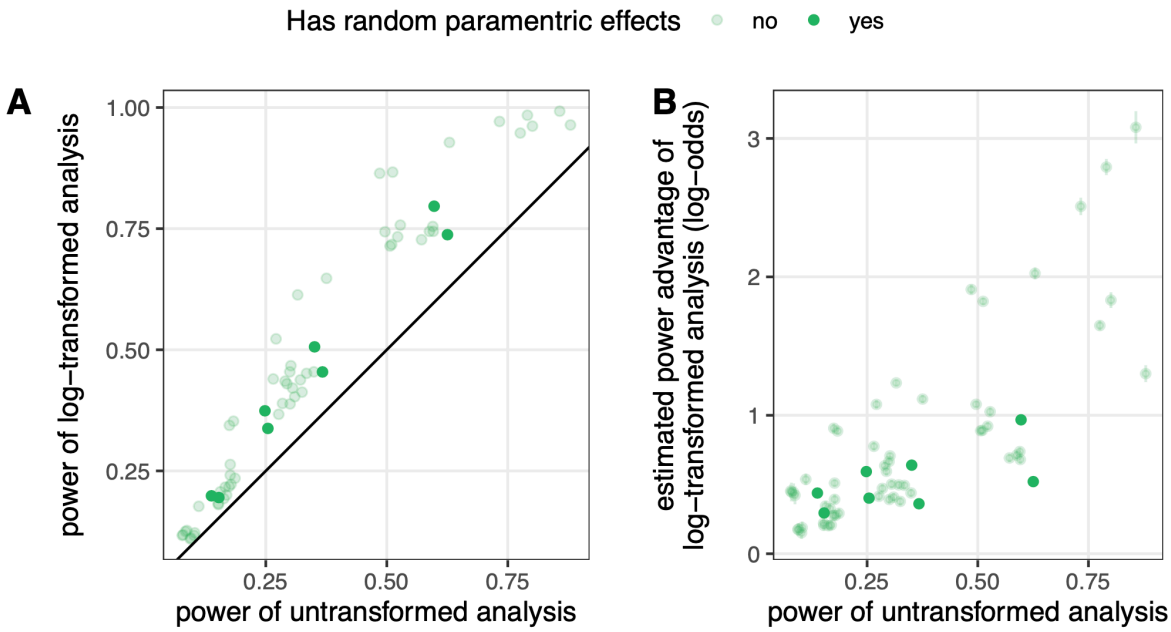

Figure 49. Same as Figure 19 in the main text, but subset to the HS18 source data, and highlighting the simulations that included parametric random effects. Comparing the Type I error-corrected power of the untransformed and log-transformed analysis approaches across all HS18 simulation conditions of Studies 2 and 2a-2e. ***Panel A:*** *power of both analysis approaches with a gray identity line.* ***Panel B:*** *power advantage (in log-odds) of the log-transformed analysis approach over the untransformed approach as a function of the untransformed analysis approach's power. Comparisons in which the log-transformed approach had 100% power are excluded. Intervals show 95% CIs from logistic regression comparing the statistical power of two approaches.*

SI-9 Study 3

Sections SI-9.1 and SI-9.2 summarize the rate of convergence and singular fits, respectively. Section SI-9.3 to SI-9.5 provide additional plots and analyses of the Type I error rates, power, and Type I error-corrected power.

SI-9.1 Convergence failures

Parametric normal data generation. The nested model comparison found that analysis approach did not have a significant effect on rate of convergence failure with the normally-distributed parametric simulations (Table 31).

Table 31: Results of nested model comparison of convergence failures in Study 3 when RTs were generated to follow a normal distribution. Comparing a full interaction model (i.e., sample size, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*
1	6	5.35e-13	1.000
2	5	6.85e-01	0.984

i	Δ_{df}	Δ_{dev}	*p*
3	1	3.34e-01	0.563

Parametric log-normal data generation. The nested model comparison found that analysis approach did not have a significant effect on rate of convergence failure with the log-normally-distributed parametric simulations (Table 32).

Table 32: Results of nested model comparison of convergence failures in Study 3 when RTs were generated to follow a log-normal distribution. Comparing a full interaction model (i.e., sample size, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*
1	6	0.145	1.000
2	5	0.158	1.000
3	1	0.037	0.847

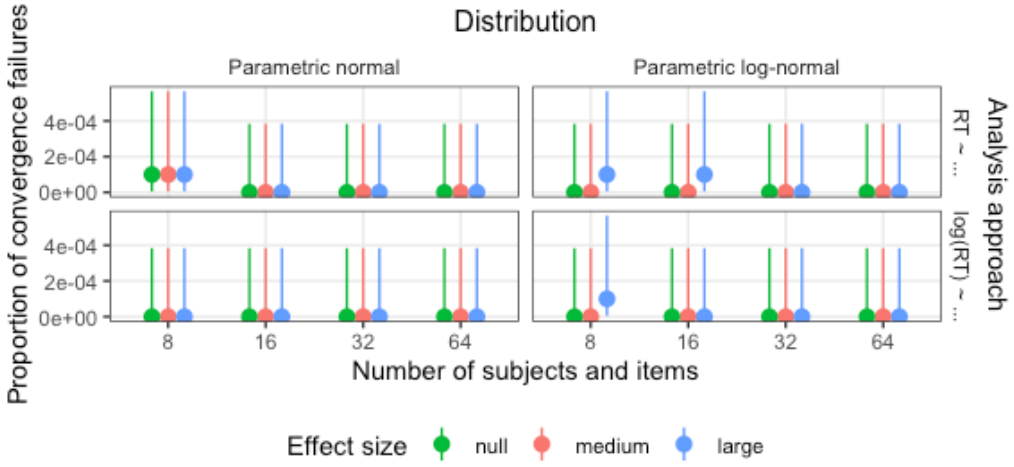

Figure 50: The convergence rate for the parametrically-generated data in Study 3. The CIs are 95% binomial score-test-based confidence intervals.

SI-9.2 Singular fits

Parametric normal data generation. The nested model comparison found that analysis approach had a significant effect on singular fits in the normally-distributed parametric simulations (Table 33).

Table 33: Results of nested model comparison of singular fit rates in Study 3 when RTs were generated to follow a normal distribution. Comparing a full interaction model (i.e., sample size, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	6	1.25	9.74e-01	
2	5	402.44	8.82e-85	***
3	1	1,075.78	6.06e-236	***

After removing the 64x64 and 32x32 sample size simulations to avoid singularities, the full interaction model found that the log-transformed analysis had significantly more singular fits than the untransformed approach ($p<0.001$). This differs from the analyses of the naturally-produced RT data, which suggested the

opposite. Analysis approach also had significant interactions with sample size (p<0.001) and effect size (p=0.015).

Parametric log-normal data generation. The nested model comparison found that analysis approach had a significant effect on singular fits in the log-normally-distributed parametric simulations (Table 34).

Table 34: Results of nested model comparison of singular fit rates in Study 3 when RTs were generated to follow a log-normal distribution. Comparing a full interaction model (i.e., sample size, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	p	
1	6	3.68	0.719540	
2	5	14.92	0.010730	*
3	1	12.75	0.000355	***

After removing the 64x64 and 32x32 sample size simulations to avoid singularities due to ceiling effects, the full interaction model found that the log-transformed analysis had significantly fewer singular fits than the untransformed approach (p<0.001). This differs from the analyses of the naturally-produced RT data, which suggested the opposite. Analysis approach also had significant interactions with sample size (p=0.013) and effect size (p=0.0054).

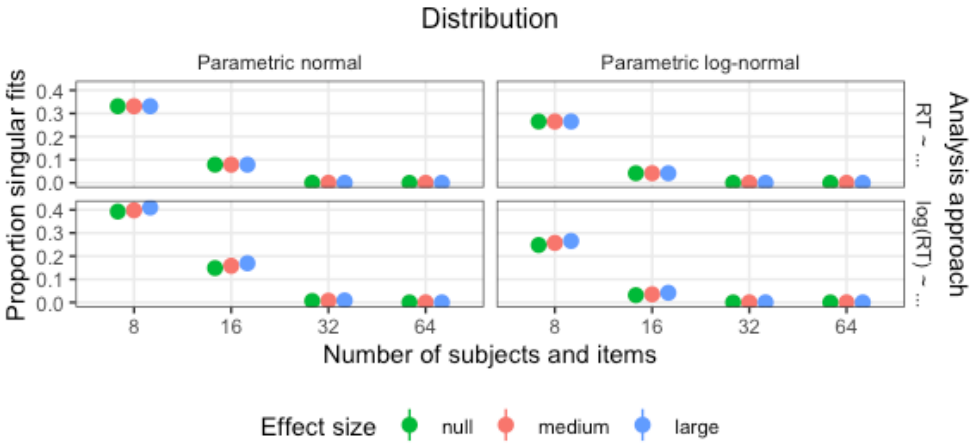

Figure 51: Proportion of singular fits for the parametrically-generated data in Study 3. 95% CIs of binomial score-test are plotted, but too small to be visible.

SI-9.3 Type I errors

Parametric normal data generation. The nested model comparison found that analysis approach did not have a significant effect on Type I error rate with the normally-distributed parametric simulations (Table 35).

Table 35: Results of nested model comparison of Type I error rates in Study 3 when RTs were generated to follow a normal distribution. Comparing a full interaction model (i.e., sample size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	p
1	3	1.21	0.7502
2	1	3.73	0.0536

Parametric log-normal data generation. The nested model comparison found that analysis approach did not have a significant effect on Type I error rate with the log-normally-distributed parametric simulations (Table 36).

Table 36: Results of nested model comparison of Type I error rates in Study 3 when RTs were generated to follow a log-normal distribution. Comparing a full interaction model (i.e., sample size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*
1	3	0.501	0.919
2	1	1.165	0.280

SI-9.4 Power (uncorrected)

Parametric normal data generation. For the normally-distributed parametric data, the log-transformed analysis approach had higher power than the standard approach for 0 of the 7 comparisons that were neither at ceiling nor at floor (of which there was 1). The log-transformed analysis approach had lower power on average, with a mean difference of 0.201 in log-odds (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of 78.3% vs. 81.6% power.

Parametric log-normal data generation. For the log-normally-distributed parametric data, the log-transformed analysis approach had higher power than the standard approach for 7 of the 7 comparisons that were neither at ceiling nor at floor (of which there was 1). The log-transformed analysis approach had higher power on average, with a mean difference of 1.01 in log-odds (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of 70.7% vs. 86.9% power.

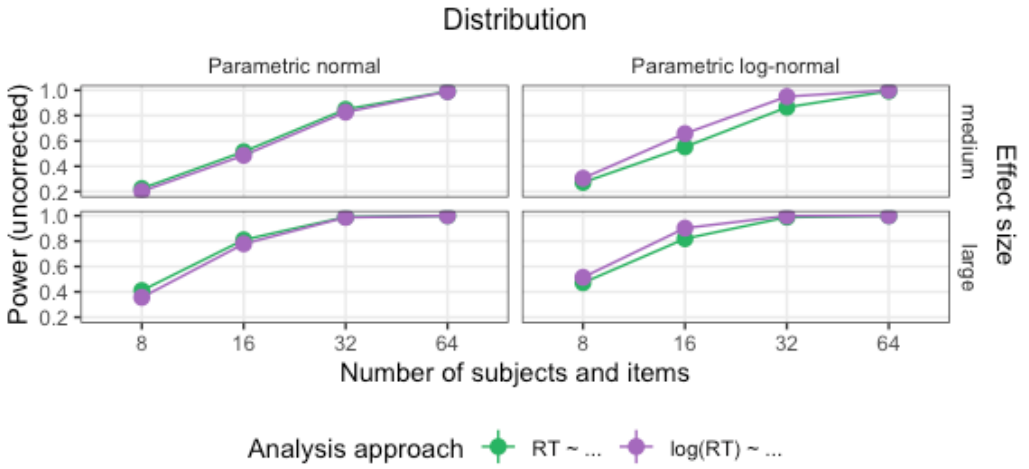

Figure 52: Power (uncorrected) for both analysis approaches on the parametric HS18 data. 95% CIs of binomial score-test are plotted, but too small to be visible.

SI-9.5 Power (corrected)

Parametric normal data generation. For the normally-distributed parametric data, the log-transformed analysis approach had lower Type I-corrected power than the untransformed approach for 7 of the 7 comparisons that were neither at ceiling nor at floor (of which there was 1). The log-transformed analysis approach had lower Type I-corrected power on average, with a mean difference of 0.15 in log-odds (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of 78.8% vs. 81.2% power. Model comparison found that analysis approach had a significant effect on Type I-corrected power (Table 37).

Table 37: Results of nested model comparison of Type I-corrected power in Study 3 when RTs were generated to follow a normal distribution. Comparing a full interaction model (i.e., sample size, effect size,

and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	3	1.28	7.33e-01	
2	4	9.04	6.00e-02	
3	1	46.95	7.29e-12	***

After removing the 64x64 sample size simulations to avoid singularities, the full interaction model found that the log-transformed analysis had significantly less corrected power than the untransformed approach (p<0.001) on the normally-distributed, parametrically-generated data. It also had a significant interaction with effect size (p=0.019), with the difference between approaches increasing for the larger effect size, and a significant interaction with sample size, with the 32x32 sample size increasing the untransformed analysis approach's advantage (p=0.083).

Table 38: The summary of the full interaction model for the Type I-corrected power in Study 3 when RTs were generated to follow a normal distribution.

term	estimate	std.error	statistic	p.value	
(Intercept)	0.962	0.0128	74.9	< 0.001	***
SampSize-8	-1.82	0.0144	-126.	< 0.001	***
SampSize-16	-0.272	0.0144	-18.9	< 0.001	***
Effect-medium	-0.840	0.0128	-65.4	< 0.001	***
Anlys-stndrd	0.0680	0.0128	5.30	1.2e-07	***
SampSize-8:Effect-medium	0.424	0.0144	29.5	< 0.001	***
SampSize-16:Effect-medium	0.164	0.0144	11.4	< 0.001	***
SampSize-8:Anlys-stndrd	-0.0220	0.0144	-1.53	0.125	
SampSize-16:Anlys-stndrd	-0.0250	0.0144	-1.73	0.083	
Effect-medium:Anlys-stndrd	-0.0302	0.0128	-2.35	0.019	*
SampSize-8:Effect-medium:Anlys-stndrd	0.00941	0.0144	0.654	0.513	
SampSize-16:Effect-medium:Anlys-stndrd	0.0155	0.0144	1.08	0.280	

***Parametric log-normal data generation*.** The log-transformed analysis approach had higher Type I-corrected power on average, with a mean difference of 0.962 in log-odds (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of 71.2% vs. 86.6% power. Model comparison found that analysis approach had a significant effect on Type I-corrected power with the log-normally-distributed parametric simulations (Table 39).

Table 39: Results of nested model comparison of Type I-corrected power in Study 3 when RTs were generated to follow a log-normal distribution. Comparing a full interaction model (i.e., sample size, effect size, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	3	34	1.95e-07	***
2	4	435	9.38e-93	***
3	1	660	1.63e-145	***

After removing the 64x64 sample size simulations to avoid singularities, the full interaction model found that the log-transformed analysis had significantly more corrected power than the untransformed approach (p<0.001) on the log-normally-distributed, parametrically-generated data. It had a marginally significant interaction with effect size (p<0.001), with the difference between approaches increasing for the larger effect size. There were a number of other significant interactions as well (see Table 40).

Table 40: The summary of the full interaction model for the Type I-corrected power in Study 3 when RTs were generated to follow a log-normal distribution.

term	estimate	std.error	statistic	p.value	
(Intercept)	1.55	0.0263	58.8	< 0.001	***
SampSize-8	-2.03	0.0270	-75.2	< 0.001	***
SampSize-16	-0.410	0.0273	-15.0	< 0.001	***
Effect-medium	-0.920	0.0263	-35.0	< 0.001	***
Anlys-stndrd	-0.393	0.0263	-14.9	< 0.001	***
SampSize-8:Effect-medium	0.482	0.0270	17.9	< 0.001	***
SampSize-16:Effect-medium	0.201	0.0273	7.36	1.9e-13	***
SampSize-8:Anlys-stndrd	0.329	0.0270	12.2	< 0.001	***
SampSize-16:Anlys-stndrd	0.105	0.0273	3.85	0.00012	***
Effect-medium:Anlys-stndrd	0.126	0.0263	4.78	1.8e-06	***
SampSize-8:Effect-medium:Anlys-stndrd	-0.123	0.0270	-4.57	4.8e-06	***
SampSize-16:Effect-medium:Anlys-stndrd	-0.0589	0.0273	-2.16	0.03102	*

SI-10 Study 4

Sections SI-10.1 and SI-10.2 summarize the rate of convergence and singular fits, respectively. Section SI-10.3 to SI-10.5 provide additional plots and analyses of the Type I error rates, power, and Type I error-corrected power.

SI-10.1 Convergence failures

In the process of analyzing the convergence failures for the BATAs for Study 4, it was observed that the convergence failure rates for the untransformed analysis approach differed significantly between interaction directions for the 8x8 BATAs, even when there were no added effects present (i.e., the data was exactly the same; Figure 53). After thoroughly investigating this unexpected difference, it appears that the only thing different between conditions was the contrast coding of the natural effect in the models, which flipped from -1 vs. 1 and 1 vs. -1 depending on the direction of the interaction. This appears to be a bug within *lme4* or the optimizer used, and the 8x8 BATAs were excluded from these analyses. This bug was filed with *lme4* via GitHub (https://github.com/lme4/lme4/issues/709).

To compare the rate of convergence failure between analysis approaches, we fit a full interaction model with sample size, the direction of the interaction effect, the effects present, the scale the effects were linear in, and analysis approach. Model comparison found that analysis approach had a significant effect on rate of convergence failure for the interaction HS18 simulations (Table 41).

Table 41: Results of nested model comparison of convergence failures in Study 4. Comparing a full interaction model (i.e., sample size, interaction effect direction, present effects, effect scale, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	6	0.478	0.9981	
2	17	2.702	1.0000	
3	17	5.597	0.9955	
4	7	10.736	0.1506	
5	1	5.109	0.0238	*

The full interaction model found that the log-transformed analysis had significantly fewer convergence failures than the untransformed approach (p=0.027), with a significant interaction with sample size (p=0.013).

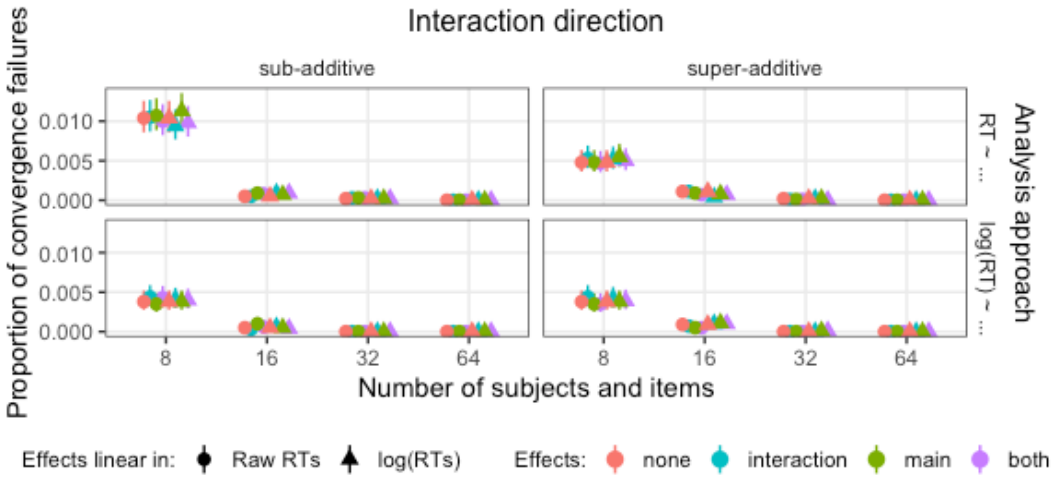

Figure 53: Convergence rates in Study 4. 95% CIs of binomial score-test are plotted, but small enough to be partially hidden.

SI-10.2 Singular fits

Model comparison found that analysis approach had a significant effect on rates of singular fit for the interaction HS18 simulations (Table 42).

Table 42: Results of nested model comparison of singular fit rates in Study 4. Comparing a full interaction model (i.e., sample size, interaction effect direction, present effects, effect scale, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	9	1.23e-02	1.00e+00	
2	24	1.02e+00	1.00e+00	
3	22	1.78e+00	1.00e+00	
4	8	2.92e+02	2.57e-58	***
5	1	8.13e+03	0.00e+00	***

The full interaction model found that the log-transformed analysis had significantly fewer convergence failures than the untransformed approach (p<0.001).

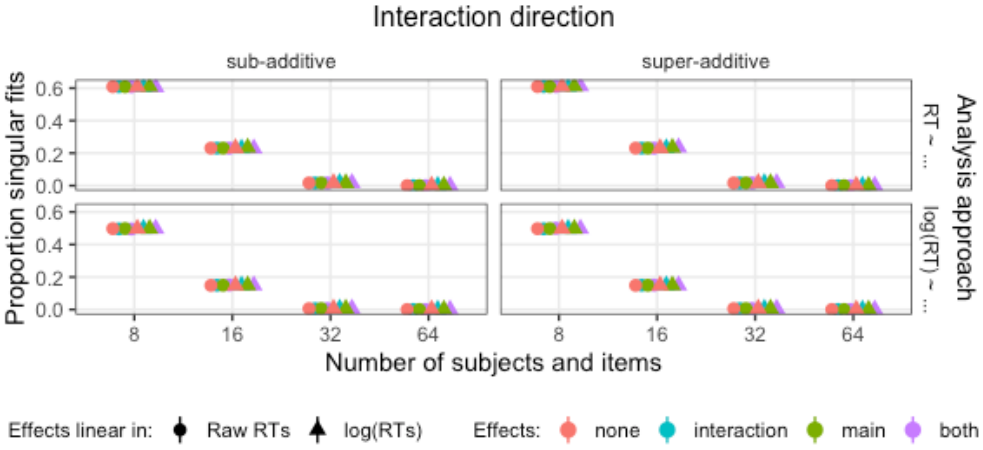

Figure 54: Rates of singular fit in Study 4. 95% CIs of binomial score-test are plotted but too small to be visible.

SI-10.3 Type I errors

We analyzed Type I error rates separately for the (smaller) main effect and the interaction.

Main effect. The nested model comparison found that analysis approach had a significant effect on Type I error rates for detecting the main effect of the interaction (Table 43). The full interaction model found that the log-transformed analysis had significantly higher Type I error rates than the untransformed approach (p<0.001). The effect of analysis approach had two interactions with sample size, with the difference between approaches being greater for the 8x8 BATAs, but less for BATAs that were 32x32.

Table 43: Results of nested model comparison of Type I error rates in Study 4 for the main effect. Comparing a full interaction model (i.e., sample size, interaction effect direction, present effects, effect scale, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	p	
1	3	0.171	0.982155	
2	10	0.590	0.999985	
3	12	0.895	0.999992	
4	6	19.455	0.003460	**
5	1	13.915	0.000191	***

Table 44: The summary of the full interaction model for Type I main effect data for Study 4.

term	estimate	std.error	statistic	p.value	
(Intercept)	-2.94	0.00573	-513.	< 0.001	***
SampSize-8	-0.0425	0.0101	-4.21	2.5e-05	***
SampSize-16	0.0223	0.00986	2.27	0.02350	*
SampSize-32	-0.0191	0.00998	-1.91	0.05607	
Dir-sub-additive	-0.00271	0.00573	-0.472	0.63674	
Effects-none	-0.00260	0.00573	-0.453	0.65034	
Scale-Raw RTs	0.000152	0.00573	0.0265	0.97883	
Anlys-stndrd	-0.0218	0.00573	-3.80	0.00015	***
SampSize-8:Dir-sub-additive	-0.00425	0.0101	-0.421	0.67350	
SampSize-16:Dir-sub-additive	-0.00102	0.00986	-0.104	0.91723	
SampSize-32:Dir-sub-additive	0.00433	0.00998	0.434	0.66406	
SampSize-8:Effects-none	0.00324	0.0101	0.322	0.74774	
SampSize-16:Effects-none	0.000386	0.00986	0.0391	0.96878	
SampSize-32:Effects-none	0.00130	0.00998	0.130	0.89621	
Dir-sub-additive:Effects-none	0.000347	0.00573	0.0606	0.95170	
SampSize-8:Scale-Raw RTs	-0.00197	0.0101	-0.196	0.84464	
SampSize-16:Scale-Raw RTs	0.00156	0.00986	0.158	0.87461	
SampSize-32:Scale-Raw RTs	0.000895	0.00998	0.0897	0.92851	
Dir-sub-additive:Scale-Raw RTs	0.00183	0.00573	0.319	0.74960	
Effects-none:Scale-Raw RTs	-0.000152	0.00573	-0.0265	0.97883	
SampSize-8:Anlys-stndrd	-0.0405	0.0101	-4.02	5.8e-05	***
SampSize-16:Anlys-stndrd	0.00332	0.00986	0.337	0.73623	
SampSize-32:Anlys-stndrd	0.0294	0.00998	2.94	0.00325	**
Dir-sub-additive:Anlys-stndrd	-0.00273	0.00573	-0.477	0.63365	
Effects-none:Anlys-stndrd	0.000447	0.00573	0.0780	0.93786	
Scale-Raw RTs:Anlys-stndrd	-0.00283	0.00573	-0.494	0.62107	
SampSize-8:Dir-sub-additive:Effects-none	-0.00257	0.0101	-0.255	0.79850	
SampSize-16:Dir-sub-additive:Effects-none	0.00312	0.00986	0.317	0.75154	
SampSize-32:Dir-sub-additive:Effects-none	-0.00197	0.00998	-0.198	0.84323	

term	estimate	std.error	statistic	p.value
SampSize-8:Dir-sub-additive:Scale-Raw RTs	-0.0000018	0.0101	-0.00017	0.99986
SampSize-16:Dir-sub-additive:Scale-Raw RTs	-0.00388	0.00986	-0.393	0.69416
SampSize-32:Dir-sub-additive:Scale-Raw RTs	0.00271	0.00998	0.272	0.78595
SampSize-8:Effects-none:Scale-Raw RTs	0.00197	0.0101	0.196	0.84464
SampSize-16:Effects-none:Scale-Raw RTs	-0.00156	0.00986	-0.158	0.87461
SampSize-32:Effects-none:Scale-Raw RTs	-0.000895	0.00998	-0.0897	0.92851
Dir-sub-additive:Effects-none:Scale-Raw RTs	-0.00183	0.00573	-0.319	0.74960
SampSize-8:Dir-sub-additive:Anlys-stndrd	-0.00242	0.0101	-0.240	0.81033
SampSize-16:Dir-sub-additive:Anlys-stndrd	0.00447	0.00986	0.453	0.65063
SampSize-32:Dir-sub-additive:Anlys-stndrd	-0.00178	0.00998	-0.178	0.85835
SampSize-8:Effects-none:Anlys-stndrd	-0.00138	0.0101	-0.137	0.89120
SampSize-16:Effects-none:Anlys-stndrd	0.00155	0.00986	0.157	0.87514
SampSize-32:Effects-none:Anlys-stndrd	0.000601	0.00998	0.0602	0.95202
Dir-sub-additive:Effects-none:Anlys-stndrd	0.00197	0.00573	0.344	0.73055
SampSize-8:Scale-Raw RTs:Anlys-stndrd	0.00277	0.0101	0.275	0.78338
SampSize-16:Scale-Raw RTs:Anlys-stndrd	0.000386	0.00986	0.0391	0.96879
SampSize-32:Scale-Raw RTs:Anlys-stndrd	0.00154	0.00998	0.154	0.87758
Dir-sub-additive:Scale-Raw RTs:Anlys-stndrd	0.000196	0.00573	0.0342	0.97274
Effects-none:Scale-Raw RTs:Anlys-stndrd	0.00283	0.00573	0.494	0.62107
SampSize-8:Dir-sub-additive:Effects-none:Scale-Raw RTs	0.00000175	0.0101	0.000174	0.99986
SampSize-16:Dir-sub-additive:Effects-none:Scale-Raw RTs	0.00388	0.00986	0.393	0.69416
SampSize-32:Dir-sub-additive:Effects-none:Scale-Raw RTs	-0.00271	0.00998	-0.272	0.78595
SampSize-8:Dir-sub-additive:Effects-none:Anlys-stndrd	0.000197	0.0101	0.0195	0.98443
SampSize-16:Dir-sub-additive:Effects-none:Anlys-stndrd	-0.00376	0.00986	-0.381	0.70296
SampSize-32:Dir-sub-additive:Effects-none:Anlys-stndrd	0.00254	0.00998	0.254	0.79915
SampSize-8:Dir-sub-additive:Scale-Raw RTs:Anlys-stndrd	-0.00279	0.0101	-0.277	0.78182
SampSize-16:Dir-sub-additive:Scale-Raw RTs:Anlys-stndrd	0.00301	0.00986	0.305	0.76006
SampSize-32:Dir-sub-additive:Scale-Raw RTs:Anlys-stndrd	-0.00180	0.00998	-0.180	0.85718
SampSize-8:Effects-none:Scale-Raw RTs:Anlys-stndrd	-0.00277	0.0101	-0.275	0.78338
SampSize-16:Effects-none:Scale-Raw RTs:Anlys-stndrd	-0.000386	0.00986	-0.0391	0.96879
SampSize-32:Effects-none:Scale-Raw RTs:Anlys-stndrd	-0.00154	0.00998	-0.154	0.87758
Dir-sub-additive:Effects-none:Scale-Raw RTs:Anlys-stndrd	-0.000196	0.00573	-0.0342	0.97274
SampSize-8:Dir-sub-additive:Effects-none:Scale-Raw RTs:Anlys-stndrd	0.00279	0.0101	0.277	0.78182
SampSize-16:Dir-sub-additive:Effects-none:Scale-Raw RTs:Anlys-stndrd	-0.00301	0.00986	-0.305	0.76006
SampSize-32:Dir-sub-additive:Effects-none:Scale-Raw RTs:Anlys-stndrd	0.00180	0.00998	0.180	0.85718

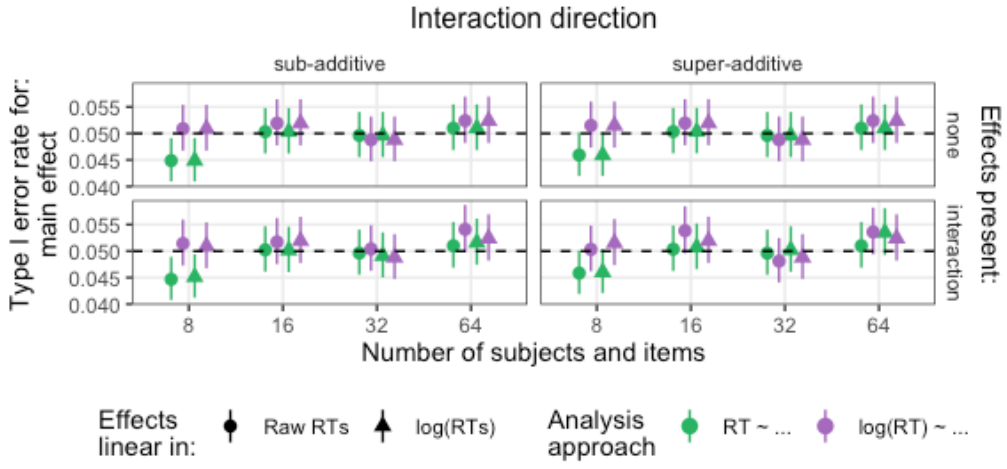

Figure 55: Type I error rates of main effect in Study 4. The CIs are 95% binomial score-test-based confidence intervals.

Interaction effect. The nested model comparison found that analysis approach had a significant effect on Type I error rate for the interaction effect (Table 45). The full interaction model found that the log-transformed analysis had significantly higher Type I error rates than the untransformed approach (p<0.001), but there were also a large number of significant higher-order interactions (see Table 46 and Figure 56).

Table 45: Results of nested model comparison of Type I error rates in Study 4 for the interaction effect. Comparing a full interaction model (i.e., sample size, present effects, effect scale, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	p	
1	3	38.5	2.21e-08	***
2	7	158.8	5.61e-31	***
3	5	150.1	1.30e-30	***
4	1	92.8	5.67e-22	***

Table 46: Summary of the full interaction model for the Type I interaction effect data in Study 4.

term	estimate	std.error	statistic	p.value	
(Intercept)	-2.95	0.00823	-358.	< 0.001	***
SampSize-8	-0.245	0.0153	-16.0	< 0.001	***
SampSize-16	-0.0199	0.0142	-1.40	0.16197	
SampSize-32	0.0900	0.0138	6.52	7.1e-11	***
Effects-none	-0.0772	0.00823	-9.38	< 0.001	***
Scale-Raw RTs	0.0171	0.00823	2.08	0.03776	*
Anlys-stndrd	-0.0770	0.00823	-9.36	< 0.001	***
SampSize-8:Effects-none	0.0538	0.0153	3.51	0.00045	***
SampSize-16:Effects-none	0.0360	0.0142	2.53	0.01136	*
SampSize-32:Effects-none	-0.0141	0.0138	-1.02	0.30835	
SampSize-8:Scale-Raw RTs	-0.0065	0.0153	-0.422	0.67332	
SampSize-16:Scale-Raw RTs	-0.0128	0.0142	-0.896	0.37007	
SampSize-32:Scale-Raw RTs	0.0120	0.0138	0.866	0.38656	
Effects-none:Scale-Raw RTs	-0.0171	0.00823	-2.08	0.03776	*
SampSize-8:Anlys-stndrd	-0.0546	0.0153	-3.57	0.00036	***
SampSize-16:Anlys-stndrd	-0.0020	0.0142	-0.143	0.88611	
SampSize-32:Anlys-stndrd	0.0237	0.0138	1.72	0.08624	
Effects-none:Anlys-stndrd	0.0171	0.00823	2.08	0.03733	*
Scale-Raw RTs:Anlys-stndrd	-0.0772	0.00823	-9.38	< 0.001	***
SampSize-8:Effects-none:Scale-Raw RTs	0.00646	0.0153	0.422	0.67332	
SampSize-16:Effects-none:Scale-Raw RTs	0.0128	0.0142	0.896	0.37007	
SampSize-32:Effects-none:Scale-Raw RTs	-0.0120	0.0138	-0.866	0.38656	
SampSize-8:Effects-none:Anlys-stndrd	-0.0065	0.0153	-0.424	0.67144	
SampSize-16:Effects-none:Anlys-stndrd	-0.0127	0.0142	-0.892	0.37254	
SampSize-32:Effects-none:Anlys-stndrd	0.0120	0.0138	0.867	0.38604	
SampSize-8:Scale-Raw RTs:Anlys-stndrd	0.0538	0.0153	3.51	0.00045	***
SampSize-16:Scale-Raw RTs:Anlys-stndrd	0.0360	0.0142	2.53	0.01136	*
SampSize-32:Scale-Raw RTs:Anlys-stndrd	-0.0141	0.0138	-1.02	0.30835	
Effects-none:Scale-Raw RTs:Anlys-stndrd	0.0772	0.00823	9.38	< 0.001	***
SampSize-8:Effects-none:Scale-Raw RTs:Anlys-stndrd	-0.0538	0.0153	-3.51	0.00045	***
SampSize-16:Effects-none:Scale-Raw RTs:Anlys-stndrd	-0.0360	0.0142	-2.53	0.01136	*
SampSize-32:Effects-none:Scale-Raw RTs:Anlys-stndrd	0.0141	0.0138	1.02	0.30835	

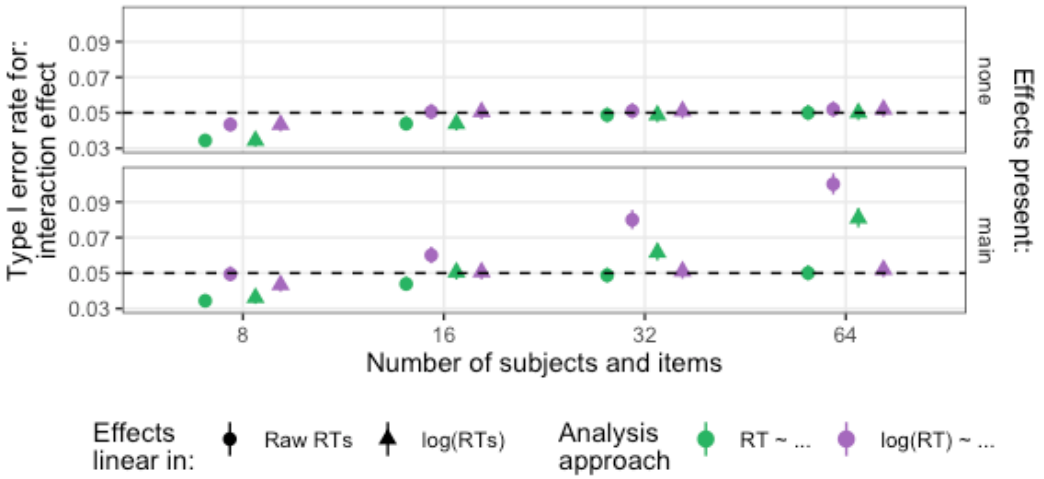

Figure 56: Type I error rates of the interaction in Study 4. 95% CIs of binomial score-test are plotted, but too small to be visible.

SI-10.4 Power (uncorrected)

We report power separately for the (smaller) main effect and the interaction.

Main effect. For the main effect of Study 4, the log-transformed analysis approach had higher power than the standard approach for 30 of the 32 comparisons that were not both at ceiling or floor. The log-transformed analysis approach had higher power on average, with a mean difference of 0.304 in log-odds (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of 77.5% vs. 82.3% power.

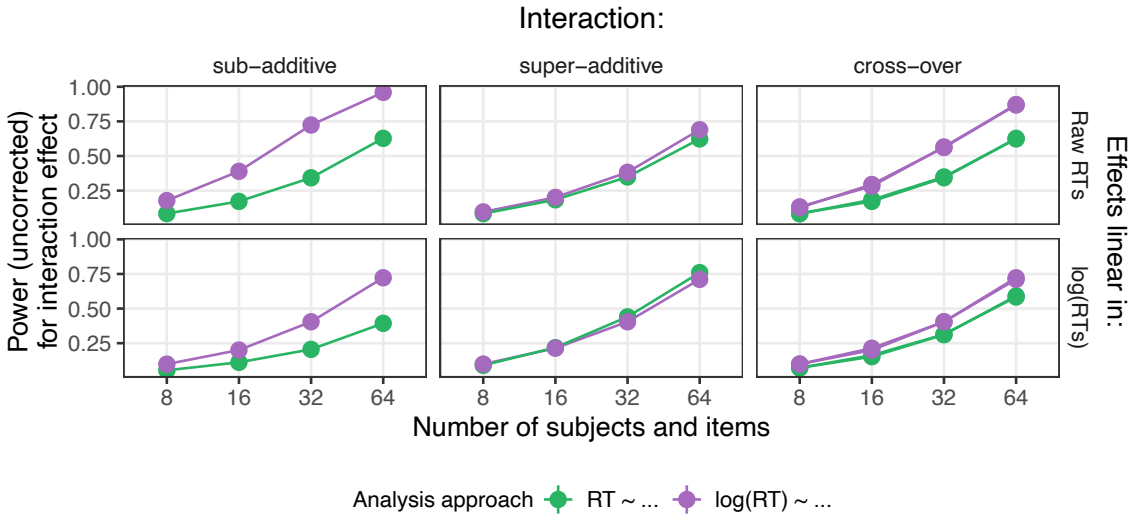

Figure 57. Power (uncorrected) for both analysis approaches of the interaction effect in Study 4. Note that an interaction effect was always present in the power simulations for the main effect in Study 4 (since Study 2 already assesses the power of the main effect in the absence of an interaction). 95% CIs of binomial score-test are plotted, but too small to be visible.

Interaction effect. For the interaction effect of Study 4, the log-transformed analysis approach had higher power than the standard approach for 29 of the 32 comparisons that were not both at ceiling or floor. The log-transformed analysis approach had higher power on average, with a mean difference of 0.638 in

log-odds (using plus-one smoothing to avoid any infinite values), which corresponds to a difference of 74.4% vs. 84.6% power.

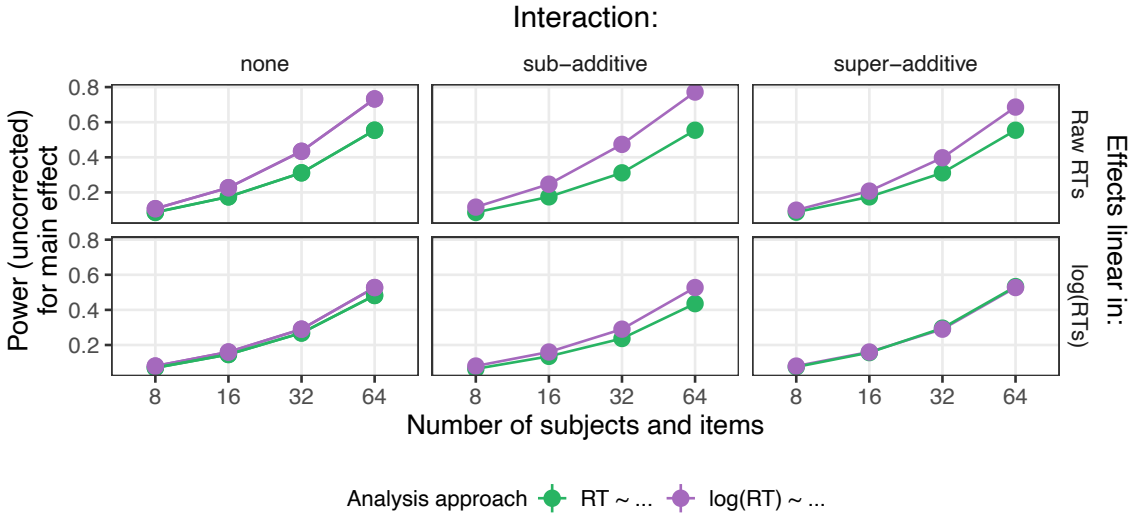

Figure 58. Power (uncorrected) for both analysis approaches of the smaller main effect in Study 4. 95% CIs of binomial score-test are plotted, but too small to be visible.

SI-10.5 Power (corrected)

We report power separately for the (smaller) main effect and the interaction.

Main effect. Model comparison found that analysis approach had a significant effect on Type I-corrected power for the main effect of the Study 4 data (Table 47).

Table 47: Results of nested model comparison of Type I-corrected power in Study 4 for the main effect. Comparing a full interaction model (i.e., sample size, interaction effect direction, present effects, effect scale, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	3	0.17	9.82e-01	
2	10	15.82	1.05e-01	
3	12	372.65	2.31e-72	***
4	6	1,622.95	0.00e+00	***
5	1	2,694.71	0.00e+00	***

The full interaction model found that the log-transformed analysis had significantly better power than the untransformed approach (p<0.001). For additional significant interactions with analysis approach, see Table 48.

Table 48: Summary of the full interaction model for Type I-corrected power in Study 4 for the main effect

term	estimate	std.error	statistic	p.value	
(Intercept)	-0.481	0.00402	-119.	< 0.001	***
SampSize-8	-1.50	0.00758	-198.	< 0.001	***
SampSize-16	-0.633	0.00636	-99.4	< 0.001	***
SampSize-32	0.375	0.00596	62.9	< 0.001	***
MeanRT-high	-0.546	0.00560	-97.6	< 0.001	***
MeanRT-medium	0.0299	0.00552	5.41	6.2e-08	***
Effect-medium	-0.500	0.00402	-124.	< 0.001	***
Anlys-stndrd	-0.227	0.00402	-56.5	< 0.001	***

term	estimate	std.error	statistic	p.value	
SampSize-8:MeanRT-high	0.201	0.0112	18.0	< 0.001	***
SampSize-16:MeanRT-high	0.146	0.00923	15.8	< 0.001	***
SampSize-32:MeanRT-high	-0.0114	0.00839	-1.36	0.17438	
SampSize-8:MeanRT-medium	0.0405	0.0105	3.86	0.00011	***
SampSize-16:MeanRT-medium	-0.0267	0.00885	-3.02	0.00255	**
SampSize-32:MeanRT-medium	-0.0258	0.00825	-3.13	0.00174	**
SampSize-8:Effect-medium	0.241	0.00758	31.8	< 0.001	***
SampSize-16:Effect-medium	0.124	0.00636	19.5	< 0.001	***
SampSize-32:Effect-medium	-0.0283	0.00596	-4.74	2.1e-06	***
MeanRT-high:Effect-medium	0.0945	0.00560	16.9	< 0.001	***
MeanRT-medium:Effect-medium	-0.00047	0.00552	-0.086	0.93180	
SampSize-8:Anlys-stndrd	0.172	0.00758	22.7	< 0.001	***
SampSize-16:Anlys-stndrd	0.0889	0.00636	14.0	< 0.001	***
SampSize-32:Anlys-stndrd	-0.0336	0.00596	-5.64	1.7e-08	***
MeanRT-high:Anlys-stndrd	0.0644	0.00560	11.5	< 0.001	***
MeanRT-medium:Anlys-stndrd	-0.00932	0.00552	-1.69	0.09124	
Effect-medium:Anlys-stndrd	0.0484	0.00402	12.0	< 0.001	***
SampSize-8:MeanRT-high:Effect-medium	-0.0533	0.0112	-4.77	1.8e-06	***
SampSize-16:MeanRT-high:Effect-medium	-0.0590	0.00923	-6.39	1.6e-10	***
SampSize-32:MeanRT-high:Effect-medium	-0.00419	0.00839	-0.500	0.61733	
SampSize-8:MeanRT-medium:Effect-medium	-0.00501	0.0105	-0.479	0.63222	
SampSize-16:MeanRT-medium:Effect-medium	-0.00170	0.00885	-0.192	0.84745	
SampSize-32:MeanRT-medium:Effect-medium	0.00582	0.00825	0.706	0.48033	
SampSize-8:MeanRT-high:Anlys-stndrd	-0.0436	0.0112	-3.90	9.5e-05	***
SampSize-16:MeanRT-high:Anlys-stndrd	-0.0232	0.00923	-2.52	0.01184	*
SampSize-32:MeanRT-high:Anlys-stndrd	0.000648	0.00839	0.0772	0.93850	
SampSize-8:MeanRT-medium:Anlys-stndrd	-0.00433	0.0105	-0.414	0.67913	
SampSize-16:MeanRT-medium:Anlys-stndrd	-0.0110	0.00885	-1.25	0.21196	
SampSize-32:MeanRT-medium:Anlys-stndrd	0.00221	0.00825	0.268	0.78902	
SampSize-8:Effect-medium:Anlys-stndrd	-0.0403	0.00758	-5.32	1.0e-07	***
SampSize-16:Effect-medium:Anlys-stndrd	-0.0286	0.00636	-4.50	6.8e-06	***
SampSize-32:Effect-medium:Anlys-stndrd	-0.00259	0.00596	-0.435	0.66335	
MeanRT-high:Effect-medium:Anlys-stndrd	-0.0161	0.00560	-2.87	0.00409	**
MeanRT-medium:Effect-medium:Anlys-stndrd	-0.00415	0.00552	-0.752	0.45218	
SampSize-8:MeanRT-high:Effect-medium:Anlys-stndrd	0.0189	0.0112	1.69	0.09088	
SampSize-16:MeanRT-high:Effect-medium:Anlys-stndrd	0.0137	0.00923	1.49	0.13615	
SampSize-32:MeanRT-high:Effect-medium:Anlys-stndrd	-0.00054	0.00839	-0.064	0.94894	
SampSize-8:MeanRT-medium:Effect-medium:Anlys-stndrd	0.00353	0.0105	0.337	0.73643	
SampSize-16:MeanRT-medium:Effect-medium:Anlys-stndrd	0.00251	0.00885	0.284	0.77638	
SampSize-32:MeanRT-medium:Effect-medium:Anlys-stndrd	-0.00171	0.00825	-0.207	0.83588	

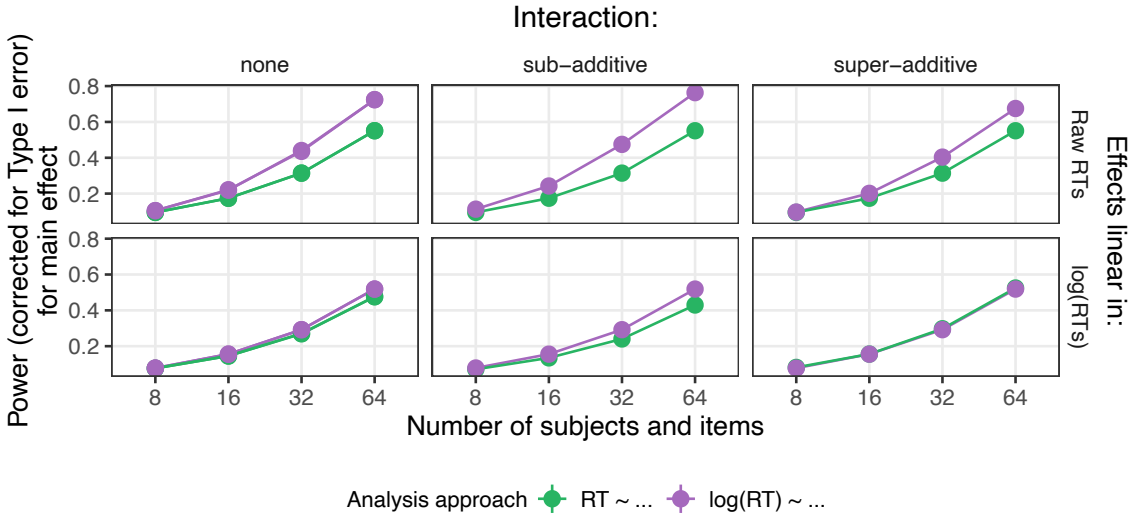

Figure 59. Power (corrected for Type I error rates) for both analysis approaches for the main effect in Study 4. Note that an interaction effect was always present in the power simulations for the main effect in Study 4 (since Study 2 already assesses the power of the main effect in the absence of an interaction). 95% CIs of binomial score-test are plotted, but too small to be visible.

Interaction effect. Model comparison found that analysis approach had a significant effect on Type I-corrected power for the interaction effect of the Study 4 data (Table 49).

Table 49: Results of nested model comparison of Type I-corrected power in Study 4 for the interaction effect. Comparing a full interaction model (i.e., sample size, interaction effect direction, present effects, effect scale, and analysis approach) to models with progressively fewer higher-order analysis approach terms (as i increases).

i	Δ_{df}	Δ_{dev}	*p*	
1	3	7.46	5.86e-02	
2	10	585.06	2.79e-119	***
3	12	3,727.12	0.00e+00	***
4	6	4,337.32	0.00e+00	***
5	1	8,407.01	0.00e+00	***

The full interaction model found that the log-transformed analysis had significantly better power than the untransformed approach (p<0.001). For additional significant interactions with analysis approach, see Table 50.

Table 50: Summary of the full interaction model for Type I-corrected power in Study 4 for the interaction effect

term	estimate	std.error	statistic	p.value	
(Intercept)	-0.481	0.00402	-119.	< 0.001	***
SampSize-8	-1.50	0.00758	-198.	< 0.001	***
SampSize-16	-0.633	0.00636	-99.4	< 0.001	***
SampSize-32	0.375	0.00596	62.9	< 0.001	***
MeanRT-high	-0.546	0.00560	-97.6	< 0.001	***
MeanRT-medium	0.0299	0.00552	5.41	6.2e-08	***
Effect-medium	-0.500	0.00402	-124.	< 0.001	***
Anlys-stndrd	-0.227	0.00402	-56.5	< 0.001	***
SampSize-8:MeanRT-high	0.201	0.0112	18.0	< 0.001	***

term	estimate	std.error	statistic	p.value	
SampSize-16:MeanRT-high	0.146	0.00923	15.8	< 0.001	***
SampSize-32:MeanRT-high	-0.0114	0.00839	-1.36	0.17438	
SampSize-8:MeanRT-medium	0.0405	0.0105	3.86	0.00011	***
SampSize-16:MeanRT-medium	-0.0267	0.00885	-3.02	0.00255	**
SampSize-32:MeanRT-medium	-0.0258	0.00825	-3.13	0.00174	**
SampSize-8:Effect-medium	0.241	0.00758	31.8	< 0.001	***
SampSize-16:Effect-medium	0.124	0.00636	19.5	< 0.001	***
SampSize-32:Effect-medium	-0.0283	0.00596	-4.74	2.1e-06	***
MeanRT-high:Effect-medium	0.0945	0.00560	16.9	< 0.001	***
MeanRT-medium:Effect-medium	-0.00047	0.00552	-0.086	0.93180	
SampSize-8:Anlys-stndrd	0.172	0.00758	22.7	< 0.001	***
SampSize-16:Anlys-stndrd	0.0889	0.00636	14.0	< 0.001	***
SampSize-32:Anlys-stndrd	-0.0336	0.00596	-5.64	1.7e-08	***
MeanRT-high:Anlys-stndrd	0.0644	0.00560	11.5	< 0.001	***
MeanRT-medium:Anlys-stndrd	-0.00932	0.00552	-1.69	0.09124	
Effect-medium:Anlys-stndrd	0.0484	0.00402	12.0	< 0.001	***
SampSize-8:MeanRT-high:Effect-medium	-0.0533	0.0112	-4.77	1.8e-06	***
SampSize-16:MeanRT-high:Effect-medium	-0.0590	0.00923	-6.39	1.6e-10	***
SampSize-32:MeanRT-high:Effect-medium	-0.00419	0.00839	-0.500	0.61733	
SampSize-8:MeanRT-medium:Effect-medium	-0.00501	0.0105	-0.479	0.63222	
SampSize-16:MeanRT-medium:Effect-medium	-0.00170	0.00885	-0.192	0.84745	
SampSize-32:MeanRT-medium:Effect-medium	0.00582	0.00825	0.706	0.48033	
SampSize-8:MeanRT-high:Anlys-stndrd	-0.0436	0.0112	-3.90	9.5e-05	***
SampSize-16:MeanRT- high:Anlys-stndrd	-0.0232	0.00923	-2.52	0.01184	*
SampSize-32:MeanRT-high:Anlys-stndrd	0.000648	0.00839	0.0772	0.93850	
SampSize-8:MeanRT-medium:Anlys-stndrd	-0.00433	0.0105	-0.414	0.67913	
SampSize-16:MeanRT-medium:Anlys-stndrd	-0.0110	0.00885	-1.25	0.21196	
SampSize-32:MeanRT-medium:Anlys-stndrd	0.00221	0.00825	0.268	0.78902	
SampSize-8:Effect-medium:Anlys-stndrd	-0.0403	0.00758	-5.32	1.0e-07	***
SampSize-16:Effect-medium:Anlys-stndrd	-0.0286	0.00636	-4.50	6.8e-06	***
SampSize-32:Effect-medium:Anlys-stndrd	-0.00259	0.00596	-0.435	0.66335	
MeanRT-high:Effect-medium:Anlys-stndrd	-0.0161	0.00560	-2.87	0.00409	**
MeanRT-medium:Effect-medium:Anlys-stndrd	-0.00415	0.00552	-0.752	0.45218	
SampSize-8:MeanRT-high:Effect-medium:Anlys-stndrd	0.0189	0.0112	1.69	0.09088	
SampSize-16:MeanRT-high:Effect-medium:Anlys-stndrd	0.0137	0.00923	1.49	0.13615	
SampSize-32:MeanRT-high:Effect-medium:Anlys-stndrd	-0.00054	0.00839	-0.064	0.94894	
SampSize-8:MeanRT-medium:Effect-medium:Anlys-stndrd	0.00353	0.0105	0.337	0.73643	
SampSize-16:MeanRT-medium:Effect-medium:Anlys-stndrd	0.00251	0.00885	0.284	0.77638	
SampSize-32:MeanRT-medium:Effect-medium:Anlys-stndrd	-0.00171	0.00825	-0.207	0.83588	

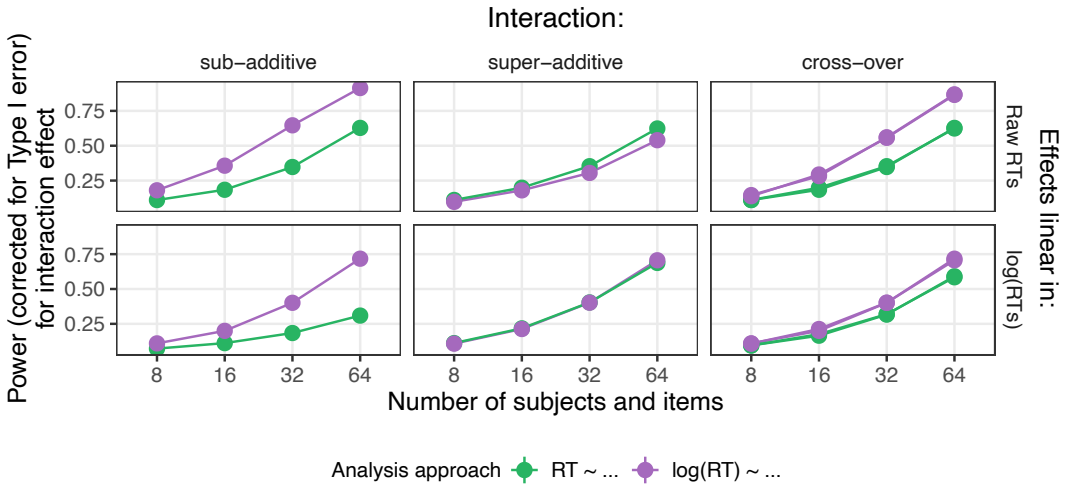

Figure 60. Power (corrected for Type I error rates) for both analysis approaches for the interaction effect of Study 4. 95% CIs of binomial score-test are plotted, but too small to be visible.

SI-11 Illustrating the effects of additive raw effects in log-RTs

Figure 61 shows how the magnitude of a (hallucinated) interaction in log-RTs depends on both the overall mean of RTs and the size of the two main effects (that are additive in raw RTs). We emphasize that Figure 61 does *not* show the results of a Type I error simulation of the type presented in Studies 2-4. It is merely demonstrating how additivity in raw RTs affects the magnitude of an interaction coefficient when the data are analyzed as log-RTs. If larger interaction coefficients always were to result in increased Type I errors, Figure 61 would also indicate expected Type I error rates. This is, however, *not* guaranteed, since RT variability also increases with larger mean RTs.

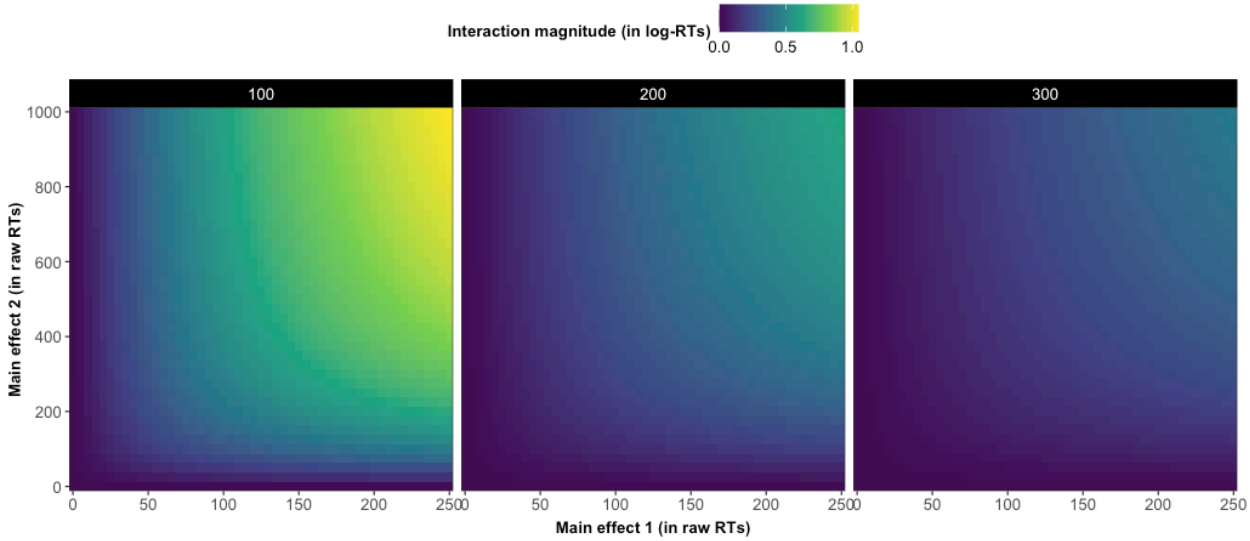

Figure 61. Illustrating how two effects that are additive in raw RTs appear as an interaction in log-RTs. Each panel shows a different RT baseline in ms (lower bound), to which the two main effects are added. The magnitude of the interaction is obtained by log-transforming the RTs of the four condition means, calculating the magnitude of simple effects of effect 2 at each of the two levels of effect 1, and then calculating the difference between the two simple effects.