AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2023, Vol. 152, No. 11, 3189–3217
https://doi.org/10.1037/xge0001450

Outlier Exclusion Procedures for Reaction Time Analysis: The Cures Are Generally Worse Than the Disease

Jeff Miller
Department of Psychology, University of Otago

A methodological problem in most reaction time (RT) tasks is that some measured RTs may be outliers, being either too fast or too slow to reflect the task-related processing of interest. Numerous ad hoc procedures have been used to identify these outliers for exclusion from further analyses, but the accuracies of these methods have not been systematically compared. The present study compared the performance of 58 different outlier exclusion procedures (OEPs) using four huge datasets of real RTs. The results suggest that these OEPs are likely to do more harm than good, because they incorrectly identify outliers, increase noise, introduce bias, and generally reduce statistical power. The results suggest that RT researchers should not automatically apply any of these OEPs to clean their RT data prior to the main analyses.

Public Significance Statement
In many areas of psychology, researchers measure people's "reaction times" (RTs), which are the times needed to perform simple decision-making tasks. Although this measure is standard, the methods for analyzing it are not, particularly with regard to the problem of identifying aberrant RTs that should be excluded from the analyses. This article evaluates a wide range of approaches to this problem so that researchers can be informed about the strengths and weaknesses of each.

Keywords: reaction time distributions, outliers, outlier exclusion procedures

Supplemental materials: https://doi.org/10.1037/xge0001450.supp

Reaction time (RT) is one of the most commonly employed dependent measures in the social sciences. It is widely used not only in many areas of psychology (e.g., cognitive, social, developmental, comparative, psycholinguistics, clinical, applied, health, etc.) but also in related disciplines such as neuroscience (e.g., Lohani et al., 2021), medicine (e.g., Baykara & Alban, 2019), sports science (e.g., Pesce & Audiffren, 2011), political science (e.g., Johnston & Madson, 2022), and economics (e.g., Clithero, 2018).

In most RT studies, researchers collect many individual-trial RTs from each participant within each of a set of experimental conditions being compared (e.g., congruent vs. incongruent flankers; Eriksen & Eriksen, 1974), generally limiting the maximum RT by having the data collection program end the trial if the participant has not responded by some maximum allowable RT. Typically, the RTs of each participant are then checked for outliers—RTs that seem spuriously fast or slow—and these outliers are discarded from further analyses. For example, many researchers compute a *Z*-score for each RT relative to the sample of all trials from the same participant and condition, and they then discard RTs with *Z*-scores whose absolute value is larger than a fixed cutoff such as 2.5 or 3.0. The rationale for eliminating these extreme RTs is that they might have been generated by anomalous processes other than the task-related

This article was published Online First July 27, 2023.
Jeff Miller https://orcid.org/0000-0003-2718-3153
The author thanks Melvin Yap and Ludovic Ferrand for their assistance in obtaining and coding the raw data from the English Lexicon Project and French Lexicon Project, respectively. The author is grateful to Alexander Berger, Patricia Haden, and Rolf Ulrich for helpful comments on earlier versions of the article. Special thanks to the University of Otago's Research and Teaching Information Technology Support Solutions division, led by Dave Robertson, and to the Department of Psychology technical staff, led by Jeremy Anderson, for assistance in providing computer facilities to run the simulations.

The author declared that he had no conflicts of interest with respect to the authorship or publication of this article.

The lexical decision task datasets were retrieved from the online repositories indicated in the original publications of Balota et al. (2007), Ferrand et al. (2010), Hutchison et al. (2013), and Keuleers et al. (2010). All computations and simulations were carried out using MATLAB. Most of the underlying general-purpose code is available at https://github.com/milleratotago/RawRT and https://github.com/milleratotago/Cupid, and other special-purpose routines are available by request from the author.

Jeff Miller served as lead for conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Jeff Miller, Department of Psychology, University of Otago, PO Box 56, Dunedin 9054, New Zealand. Email: miller@psy.otago.ac.nz

processes in which the researcher is interested, possibly due to stimulus anticipation, distraction from the experimental task, or equipment malfunction. Naturally, to get the most accurate information about the task-related processes under study, researchers would like to limit the analysis to valid RTs that were actually generated by the processes of interest. Doing so is not straightforward, however, because these processes may themselves occasionally generate quite fast or slow valid RTs. Thus, there is no guarantee that an RT which appears unusual with respect to the other RTs was necessarily perturbed by anomalous processes (Kimber, 1990). In fact, extreme-yet-valid RTs could be very revealing about the processes under study if, for example, an experimental manipulation produced a large effect on a small proportion of the trials (e.g., Tukey, 1959).[1]

In practice, RT researchers use a variety of ad hoc outlier exclusion procedures (OEPs) with differing and somewhat arbitrary cutoffs (Leys et al., 2013; Simmons et al., 2011), as will be described in the next section. Although each of these procedures has a plausible intuitive rationale, it is far from clear which one works most accurately in practice, which may be why researchers rarely state why they chose one specific procedure or cutoff in preference to the others. Given that researchers have little clear basis on which to choose between OEPs and that different OEPs do produce different final results, it is virtually certain that researchers exclude outliers suboptimally, with unknown consequences for statistical power and effect-size estimation (cf., Ratcliff, 1993; Zhou & Krott, 2016). Moreover, variation across labs in OEPs is problematic because it contributes to effect-size heterogeneity and reduces replicability when the effect size depends on the OEP (Kenny & Judd, 2019), as the present simulations will show that it does (e.g., Figure 15b). Standardization is an integral part of the scientific method (e.g., Paylor, 2009), and that surely includes methods for analyzing data as well as for collecting it. Of course, standardization using the best method would be ideal.

A more insidious problem with the lack of standardized procedures and cutoffs for outlier exclusion is that having a choice among different options provides "researcher degrees of freedom" in the analysis process, which can inflate Type 1 error rates (Leys et al., 2019; Simmons et al., 2011). A researcher might, for example, perform an initial analysis with one procedure or cutoff, obtain a near-significant result (e.g., $p = .06$), and then reanalyze the data with a different procedure or cutoff in the hopes of obtaining stronger results with "cleaner" data (i.e., $p < .05$). Such practices increase the likelihood of Type 1 errors, as has been clearly demonstrated in simulations of analyses using multiple OEPs (Bakker & Wicherts, 2014; Berger & Kiefer, 2021; Morís Fernández & Vadillo, 2020; Ulrich & Miller, 2020).

In order to work toward a better understanding of the strengths and weaknesses of alternative OEPs and hopefully promote standardization of this important data-precleaning step, the present article reports simulations comparing several popular OEPs in detail. Ideally, one would like to find a single procedure that would be most effective at identifying outliers, would yield the greatest statistical power, and would produce the most accurate effect-size estimates. Simulations with this goal have also been reported previously, of course, especially those of Berger and Kiefer (2021), Bush et al. (1993), Ratcliff (1993), Ulrich and Miller (1994), and Van Selst and Jolicœur (1994). And yet, the remaining between- and even within-lab variability in OEP usage attests to the lack of clear evidence about which procedure works best.

The present simulations extended previous work in three primary ways. First, they investigated a wider range of OEPs than have been examined in any previous study. Second, they assessed the consequences of each OEP with respect to a wide range of outcome measures—including some at the level of RT distributions. Existing studies have focussed primarily on specific consequences concerning mean RTs or experimental power. Third, the present analyses were based on actual RT datasets. This allows OEP performance to be evaluated under conditions known to be realistic with respect to the underlying distributions of valid RTs, the underlying distributions and proportions of fast and slow outliers, and the underlying effects of the experimental manipulations. Some of these analyses were applied directly to the actual observed RTs themselves, others were applied to simulations using random subsamples of the real datasets, and others were applied to artificial datasets modeled as closely as possible on the observed RT datasets. In contrast, previous simulations have all started with assumptions about the distributions of valid RTs, the distributions of fast and slow outlier RTs, the probabilities of those two types of outliers, and the distribution-level effects of the experimental manipulations (e.g., did they simply shift the valid RT distribution by a constant or did they instead change the shape of that distribution?). Unfortunately, these assumptions were necessarily somewhat speculative and arbitrary, because "evidence about the precise outlier distribution is lacking" (Berger & Kiefer, 2021, p. 12). Thus, it is useful to supplement these assumption-based simulations with analyses using actual RT data that are realistic by definition.

It might seem impossible to evaluate OEPs using real data because—unlike in assumption-based simulations—it is impossible to have a priori knowledge of which real RTs are valid and which are outliers. However, the present article introduces several novel metrics that provide insight into the relative utility and accuracy of OEPs without requiring any such assumptions. Each of these metrics is somewhat indirect, but the picture that emerges from considering all of them together is that none of the OEPs currently available actually work very well. My conclusion is that RT researchers should normally analyze all of the RTs as recorded, throwing out only rare RTs that cannot possibly be valid for the task under study (e.g., <150 ms in a choice RT task).

The organization of this article is as follows: The next two sections describe the different OEPs that were examined and the different real datasets to which these procedures were applied. Subsequent sections then describe the results obtained when applying the different OEPs to these datasets and to randomly selected subsamples of them (e.g., as needed to examine power).

Outlier Exclusion Procedures

Table 1 lists the 13 major classes of OEPs examined in this article. These were selected to cover a very wide range of OEPs that have been considered—and, in most cases, have been used—for

[1] An anonymous reviewer pointed out that the converse could also be true—that is, in some trials anomalous processes could produce ordinary-looking RTs, which should in theory also be classified as outliers. It seems impossible to identify any such ordinary-looking RTs as outliers, however, except perhaps using ancillary dependent measures such as eye movements or EEGs, so such RTs must simply be accepted as one of the contributors to random error. Since they do not have extreme values, they would presumably not have a great impact on the overall results.

Table 1
Exclusion Procedures for Identifying RT Outliers

Type of procedure	Cutoffs
No exclusions	n.a.
Fixed cutoffs	200–2,000, 200–2,500, 200–3,000
Z of RT	2, 2.5, 3
Nonrecursive Z	Depends on N trials
Simple recursive Z	2, 2.5, 3
Modified recursive Z	2, 2.5, 3
Z of log-transformed RT	2, 2.5, 3
Z of $\sqrt{}$ transformed RT	2, 2.5, 3
Median absolute deviation	2, 2.5, 3
Trimmed	1%, 2%, 5%
Tukey fences	1.5, 3
Tukey fences (skewed)	1.5, 3
Ueda's U_t	n.a.

Note. Considering the different cutoff values as well as procedures, this table exhibits 31 different exclusion procedures that could be used to identify RT outliers. RT = reaction time.

removing outliers from RT data.[2] As is described in detail below, most are applied with researcher-selected exclusion cutoffs determining how extreme an RT must be before it is classified as an outlier; the table also lists the cutoffs examined with these procedures in the current study. For convenience, different combinations of procedure-plus-cutoff will generally be referred to with the procedure type and cutoff value (e.g., "simple recursive 2.5" or "$Z_{\sqrt{}}$ 2"). Note that a "no exclusion" procedure is included as the first option in the table. This refers to the procedure of simply analyzing all RTs, and it is a useful baseline procedure against which the various OEPs can be compared.

Fixed Cutoff OEPs

Perhaps the simplest type of OEP is to use fixed absolute cutoffs to determine which RTs are inliers versus outliers—a procedure whose consequences were studied in detail by Ulrich and Miller (1994). After visually examining a distribution of observed RTs pooled across all participants and conditions, for example, a researcher might choose a range of acceptable RTs such as 200–2,000 ms, excluding all RTs outside that range as outliers (e.g., Hübner et al., 2019; Miller & Tang, 2021; Myers et al., 2022).

Adaptive OEPs

Instead of using fixed cutoffs, researchers quite commonly use one of the remaining *adaptive* OEPs in Table 1. With these OEPs, the RT exclusion cutoffs for a given participant are computed from the participant's observed RTs rather than being prespecified. The rationale for adaptive procedures is that they can adjust for the inevitable overall RT differences between participants and conditions. For example, it seems quite reasonable to exclude an RT of 1,200 ms as an outlier if the participant's other RTs in the same condition vary from 500 to 900 ms but to accept that same RT as an inlier if their other RTs range from 800 to 1,190. Obviously, an OEP must be adaptive for outlier classifications to depend on the set of observed RTs.

One very simple adaptive OEP is to use cutoffs based on the Z-scores of each participant's observed RTs (e.g., Janczyk & Ulrich, 2019), and this may be the most common approach (Leys et al., 2013). With this procedure, the mean and standard deviation of a participant's RTs are first computed. Then, a Z-score is computed for each RT based on that mean and standard deviation, and an RT is excluded as an outlier if its Z-score is sufficiently extreme (e.g., cutoffs of ± 2, ± 2.5, or ± 3).

Van Selst and Jolicœur (1994) suggested three variants of Z-score-based procedures that have also been popular (e.g., Brosowsky & Egner, 2021; Cochrane & Pratt, 2022). In one—the "nonrecursive moving Z" variant—the Z-score cutoffs are adjusted based on the number of trials in a condition so as to minimize bias, using predetermined cutoffs based on their computer simulations. Specifically, the cutoff is $Z = \pm 2.5$ when there are at least 100 trials, but the cutoff decreases as the number of trials decreases, down to a minimum of $Z = \pm 1.458$ with four trials (Van Selst & Jolicœur, 1994; Table 4). A second variant—the "simple recursive Z" procedure—uses Z-score cutoffs in the first step. If any RTs are excluded as outliers at this step, however, new Z-scores are computed using the mean and standard of the remaining RTs, and further RTs are excluded if the new Z-scores are too extreme. This process is repeated until no further RTs are identified as outliers. The third variant—the "modified recursive Z" procedure—is similar, but it is based on the rationale that the outliers should be excluded *before* the computation of the mean and standard deviation used to determine Z-scores, so that the outliers will not contaminate these sample statistics. Thus, with this procedure, the most extreme fast and slow RTs are temporarily excluded from the sample before computing the sample mean and standard deviation. Then, Z-scores are computed for all RTs, and RTs with too-extreme Z-scores are excluded as outliers. As in the simple recursive procedure, this process is repeated until no further outliers are identified.

Other researchers have used alternative approaches in which Z-scores are computed from *transformed* RT values, with the transformation designed to produce more symmetrical distributions. For example, RTs can be log-transformed before computation of Z-scores, with outliers identified based on the resulting $Z_{\log}$ values (e.g., Seow & Fleming, 2019; Wang et al., 2022). Cousineau and Chartier (2010) examined several such approaches and favored a square-root transformation in which each individual-trial RT_i is transformed with

$$y_i = \sqrt{\frac{\mathrm{RT}_i - \min(\mathrm{RT})}{\max(\mathrm{RT}) - \min(\mathrm{RT})}}. \quad (1)$$

Subsequently, a Z-score $Z_{\sqrt{},i}$ is computed for each RT_i from its transformed value with

$$Z_{\sqrt{},i} = \frac{y_i - \bar{y}}{s_y}, \quad (2)$$

[2] This table is not intended to be a complete list of all possible procedures used in RT analysis, and the vast range of such procedures would make it virtually impossible to compile one. Other procedures, for example, may attempt to minimize the effects of outliers by using other summary statistics (e.g., medians) or by basing outlier identification on specific process models of RT (e.g., Alexandrowicz, 2020). Researchers may also successively apply more than one OEP, such as first excluding RTs outside of a fixed range with an upper cutoff less than the maximum RT allowed by the data collection program (e.g., 200–3,000 ms), and then excluding RTs from the remaining subset using a Z-score cutoff (e.g., Yap et al., 2012).

where $\bar{y}$ and s_y are the mean and standard deviation of the y_i values. RT_i is rejected as an outlier if the absolute value of its corresponding $Z_{\sqrt{},i}$ exceeds a cutoff (e.g., 2.5).

An alternative to exclusions based on *Z*-scores is to exclude RTs based on their absolute deviation from the median RT (e.g., Leys et al., 2013). This approach has the advantage over *Z*-score approaches that both the median and the median absolute deviation—on which exclusions are based—are less sensitive to outliers than the sample mean and standard deviation. Formally, let RT_i, $i = 1 \ldots n$, be a sample of observed RTs and let M_T be the median of these RTs. One then converts the series of RTs to absolute deviation values $D_i = \text{abs}(RT_i - M_T)$, finds the median M_D of the D_i values, and computes the median absolute deviation (MAD) as $\text{MAD} = 1.4826 \times M_D$, with the scaling constant of 1.4826 based on an assumption of approximate normality (e.g., Leys et al., 2013). Then, each RT_i's scaled absolute distance from the median is computed as

$$\text{MAD}_i = \frac{D_i - M_D}{\text{MAD}}, \tag{3}$$

and RT_i is rejected if MAD_i is larger than a cutoff.

Trimming is another adaptive procedure for excluding outliers. With this procedure, a certain percentage of the scores are eliminated at each extreme of the sample (e.g., exclude the fastest and slowest 5% of RTs). A complication with trimming a small percentage of scores is that for certain sample sizes, the to-be-trimmed percentage is not a whole number. For example, 5% of a sample of 30 is 1.5, so it is impossible to exclude exactly the desired number of scores. In the present analyses, trimming was carried out by rounding the desired number of to-be-excluded RTs downwards for computing the number of fast outliers to be excluded and rounding upwards for computing the number of slow outliers to be excluded.

Tukey (1977) introduced another well-known procedure for identifying outliers based on a sample's interquartile range (IQR). Applying this procedure to a sample of RTs, the IQR is first computed as the distance between the RTs at the first and third quartiles, $\text{IQR} = q_3 - q_1$. Then, upper and lower "fences" are computed as $F_u = q_3 + 1.5 \times \text{IQR}$ and $F_l = q_1 - 1.5 \times \text{IQR}$, and RTs outside these fences are rejected as outliers. Clearly, the constant multiplier 1.5 used here to compute the fence boundaries is somewhat arbitrary, and larger values such as 3.0 are sometimes used (e.g., Keuleers et al., 2012), naturally excluding fewer RTs as outliers.

Tukey's fences do not seem ideal for skewed distributions because the tails of these distributions can extend different amounts below q_1 versus above q_3 (e.g., Aucremanne et al., 2004; Kimber, 1990). A possible improvement with skewed distributions is to determine the fence locations using the lower and upper semi-IQR (SIQR) assessed relative to the median (q_2), $\text{SIQR}_l = q_2 - q_1$ and $\text{SIQR}_u = q_3 - q_2$, which tend to differ from each other for skewed distributions (e.g., Hubert & Vandervieren, 2008). The upper and lower cutoffs for skew-sensitive fences could then be computed as[3] $F_{s,u} = q_3 + 2 \times 1.5 \times \text{SIQR}_u$ and $F_{s,l} = q_1 - 2 \times 1.5 \times \text{SIQR}_l$. Thus, a cutoff would be located farther from its quartile when its SIQR value was larger due to skew. Given that RT distributions are typically skewed, it is worthwhile to examine the effect of using skew-sensitive fences as well as the standard symmetric Tukey fences.[4]

Finally, Marmolejo-Ramos et al. (2015) suggested that Ueda's 1996/2009 OEP based on the Akaike Information Criterion (AIC) might be useful with skewed distributions such as RTs (also see Shimizu, 2022). In essence, the procedure involves the somewhat elaborate computation of a U_t value, which is analogous to an AIC, for each of a series of models differing only according to which RTs are assumed to be outliers. For example, one U_t is computed under the assumption that there are no outliers, another is computed assuming that just the largest RT is an outlier, another is computed assuming that just the two largest RTs are outliers, and so on. The best-fitting model is the one with the lowest U_t value, and the scores identified as outliers within this model are thus classified as the actual outliers.[5]

Condition-Specific Versus Pooled Application of Adaptive OEPs

A further complication regarding the adaptive OEPs is that they can be applied using either of two different approaches that could be called "condition-specific" versus "pooled".[6] With the condition-specific approach, the OEP is applied separately to the RTs of each participant *in each condition* (e.g., Brosowsky & Egner, 2021; Cochrane & Pratt, 2022; Janczyk & Ulrich, 2019; Plater et al., 2020; Van Selst & Jolicœur, 1994), so that outliers are identified as RTs that are discrepant from a participant's other RTs in that same condition. For example, the *Z*-score for each RT in a given condition would be computed using the mean and standard deviation of the participant's RTs only within that condition. This approach is attractive because it seems reasonable to judge whether an RT is aberrant by comparing it to other supposedly equivalent RTs. If there are mean RT differences between conditions, these should presumably be taken into account just like mean RT differences between participants.

Alternatively, with the pooled approach, the OEP is applied to the participant's RTs pooled together across all conditions (e.g., Ratcliff, 1993; Wang et al., 2022). For example, each RT's *Z*-score would be computed based on the mean and standard deviation of all of the participant's RTs regardless of condition. André (2022) argued for this approach on the grounds that the data treatment method should logically be blind to the experimental hypothesis, and he presented simulations showing that condition-specific OEP application can cause Type 1 error rates to be inflated under certain conditions (but see Karch, 2023).

Intuitively, it seems likely that any given OEP would produce approximately the same results with condition-specific versus pooled application because the same computational procedure is being applied to similar sets of RTs. Two differences can be predicted on theoretical grounds, however. First, the adaptive cutoffs obtained with pooled application should be more stable than those

[3] The multiplicative constants used with SIQR are approximately double those used with F_u and F_l to compensate for the fact that each SIQR is only approximately half of the IQR.

[4] Even more complicated and computationally intensive methods of computing fences for skewed distributions have also been proposed (e.g., Hubert & Vandervieren, 2008), but these were not included in the present article because they do not appear to have been used in RT analysis.

[5] For a more detailed explanation of the procedure, see Marmolejo-Ramos et al. (2015). Also, note that there is an error in Equation (2) of Ueda 1996/2009, where $-\sqrt{2}$ should be $+\sqrt{2}$.

[6] Unfortunately, it is not always clear from published reports which of these two approaches was actually used. I thank Alexander Berger for suggesting that the pooled approach should be included in the present analysis.

obtained with condition-specific application. The former's cutoffs are based on larger numbers of trials (i.e., combined across conditions), so there would be smaller random fluctuations in the sample statistics on which these cutoffs are based (e.g., sample mean and standard deviation for *Z*-score cutoffs). Second, the pooled approach tends to artificially underestimate the true size of the effect under study. This happens because it tends to exclude more fast RTs from the fast condition and more slow RTs from the slow one. With two conditions, for example, if a researcher excludes the fastest and slowest 2% of trials from each participant, the excluded fast RTs would tend to come from the faster condition and the excluded slow RTs would tend to come from the slower one. This increases the estimated mean RT in the faster condition (because its especially fast RTs have been excluded) and reduces the estimated mean RT in the slower condition (because its especially slow RTs have been excluded). Overall, then, the estimated experimental effect after exclusion will tend to be smaller than the true effect—that is, the true effect will be underestimated. This source of underestimation bias is not present with the condition-specific approach, because that approach considers each condition's fast and slow RTs separately.

It is impossible to specify the sizes of these two predicted differences between the condition-specific and pooled approaches or their influences on key properties of statistical tests such as power and Type 1 error rate. Because both approaches are currently in use, it is important to assess them both via simulations to find out how each might affect researchers' overall conclusions with real data. In summary, then, the present simulations examined the performance of 58 different OEPs: the four nonadaptive OEPs shown in the first two lines of Table 1, plus the 27 adaptive OEPs shown in the remaining lines of the table applied in a condition-specific manner, plus the same 27 adaptive OEPs applied in a pooled manner.

Datasets

The performance of the different OEPs in Table 1 was investigated using the four lexical decision task RT datasets summarized in Table 2: the Semantic Priming Project (SPP) of Hutchison et al. (2013), the Dutch Lexicon Project (DLP) of Keuleers et al. (2010), the English Lexicon Project (ELP) of Balota et al. (2007), and the French Lexicon Project (FLP) of Ferrand et al. (2010). In each trial of these lexical decision tasks, participants saw a visually presented letter string and made a manual response to indicate whether it was a word or nonword within some maximum time interval allowed for responding, as is usual in RT experiments. These four datasets were selected because they are very large, with many participants, many trials per participant in each condition (i.e., words vs. nonwords), or many of both. As actual RTs from participants in peer-reviewed experiments, these datasets thus provide high-resolution information about realistic RT distributions that can be used to examine the effectiveness of the different OEPs. As will be seen later, the presence of stimulus variability within the datasets (e.g., due to word frequency) also provides a helpful guide as to which responses are quite fast or slow because they are outliers and which are fast or slow because the stimulus items are easy or difficult.

For the present purposes, it is both an advantage and a disadvantage that the same task was used in all four studies. The lexical decision task is a widely used and typical RT task, so these datasets should be reasonable representations of RT distributions in general. The focus on datasets from a single task does perhaps limit the generality of the present analyses to some degree because other tasks may produce differently shaped RT distributions and thus their results may be affected differently by the various OEPs. On the other hand, there is no clear reason to expect that widely discrepant outlier RTs due to such things as anticipations and distraction would differ much from one task to another. Furthermore, focusing on a single task also allows us to estimate the likely importance of lab-to-lab variations associated with different equipment, data collection procedures, participant populations, etc.

Although each of these studies used a lexical decision task and found that average RTs were smaller to words than to nonwords, three notable differences among the studies are evident in Table 2. One is that each participant was tested in a much larger number of trials in the DLP study than in the other three studies, so these participants were much more practiced at the task. A second difference is that the size of the word/nonword effect varied substantially across datasets. The effect was much smaller in the DLP dataset, possibly because the participants were more highly practiced, or perhaps it was the case that the nonwords were easier. The third difference is that participants in different studies were allowed different maximum amounts of time to respond, ranging from 2 to 4 s. This experimenter-determined maximum allowable RT clearly limits the range of slow RT outliers that could possibly be observed. For example, approximately 0.5% and 2.0% of RTs were greater than 2 s in the SPP and ELP datasets, respectively, but it is unknown whether the procedural difference allowed for the recording of more slow outliers or more valid-but-slow RTs in these studies.

Table 2
Datasets Used to Compare OEPs

Property	Dataset			
	Semantic Priming Project	Dutch Lexicon Project	English Lexicon Project	French Lexicon Project
Citation	Hutchison et al. (2013)	Keuleers et al. (2010)	Balota et al. (2007)	Ferrand et al. (2010)
N participants	503	39	791	914
N trials	796,519	978,935	2,340,439	1,682,137
Max (RT)	2,999	1,999	4,000	1,999
Word RT	681.3	646.5	793.5	749.7
Nonword RT	752.7	676.3	881.2	846.3
Effect size	71.4	29.8	87.7	96.6

Note. The number of participants, number of trials, and mean RT values are those obtained after data cleaning. OEP performance was compared using these four datasets, four transposed versions of them, and four versions of them constructed to be free of outliers; see text for details. OEP = outlier exclusion procedure; RT = reaction time.

It is an open question how the different OEPs might be sensitive to this aspect of the procedure, and it is possible that OEPs would have greater value in studies with much larger limits—or none at all—on the maximum allowable RT.

Data Cleaning

Preliminary examination of the original datasets indicated that they contained some unrealistically small RTs, including many trials with recorded RTs less than 10 ms. It seemed inappropriate to include these obvious outliers in the analysis, since they could interfere with the performance of adaptive OEPs (e.g., by inflating the standard deviations involved in computing *Z*-scores). The data were therefore cleaned before any analyses, excluding all trials with RTs less than 150 ms. Although this cutoff is admittedly somewhat arbitrary, it seems reasonable because it is near the estimated physiological lower limit for a simple RT to a visual stimulus (Donders, 1868/1969; Luce, 1986), and because in the present datasets responses below this limit were at essentially chance accuracy. Fast outliers could still be identified after this data cleaning, of course; for example, a *Z*-score procedure would classify an RT of 200 ms as an outlier for a participant whose mean and standard deviation of RT were 800 and 100 ms, respectively.

Percentages of Outliers Excluded

One way to compare OEPs is by looking at the percentages of trials they exclude as outliers. RT researchers typically exclude a small percentage of trials as outliers—approximately in the range of 1%–5%—and a reasonable first step is to see whether each of the OEPs in fact excludes approximately this expected percentage of trials. For each of the four original datasets (i.e., SPP, DLP, ELP, and FLP), Figure 1a–1d shows the percentages of trials excluded as fast and slow outliers by each of the different OEPs using both condition-specific and pooled OEP applications.

Figure 1 shows that there are clear differences among OEPs that are consistent across datasets. Figure 1a and 1c shows that none of the OEPs excluded many fast outliers (i.e., small RTs), whereas Figure 1b and 1d shows that almost all excluded noticeably more RTs as slow outliers (large RTs). This fast/slow exclusion asymmetry is to be expected for many OEPs because of the long tails at the high ends of RT distributions. Notably, the different OEPs differed substantially with respect to the proportions of slow outliers excluded. Although most excluded fewer than 5% of RTs as slow outliers, as expected, several excluded approximately 10%, and the recursive procedures of Van Selst and Jolicœur (1994) excluded more than 25% when used with the lowest *Z*-score cutoff (i.e., 2), which is surely excessive.

Not surprisingly, Figure 1 shows that condition-specific and pooled application of the adaptive OEPs exclude similar percentages of outliers overall, but Figure 2 shows that the two approaches differ with respect to the exclusion of RTs from the fast and slow conditions (in this case, word vs. nonword). Each point in this figure corresponds to one of the 27 adaptive OEPs, and it is plotted in terms of its probability of excluding a word (horizontal axis) or a nonword (vertical axis) as an outlier. With condition-specific OEP application, approximately the same percentages of trials are excluded for words and nonwords (i.e., open circles on or near the diagonals), and this is true for both fast and slow outliers. As predicted earlier on theoretical grounds, however, with pooled application more fast outliers are excluded for the faster condition than for the slower condition (i.e., words vs. nonwords; filled circles below the diagonal), whereas more slow outliers are excluded for the slower condition than for the faster one (i.e., nonwords vs. words; filled circles above the diagonal). This dependence of outlier exclusions on word/nonword differences with pooled OEP application is evident in each of the four datasets, and it arises simply because responses are faster with words than nonwords. If outliers are truly generated by task-unrelated processes, however, they should be randomly distributed across conditions. Thus, this condition dependence of outlier exclusion for the pooled approach must be regarded as undesirable. Ironically, even though pooled OEP application is blind to the experimental condition, the approach ends up treating the two conditions asymmetrically because of the condition differences in mean RT. The asymmetric effects found with pooled OEP application have important consequences for effect-size estimation, as will be seen later (Figure 15b).

Exclusions from Zero-Outlier Datasets

The results shown in Figure 1 come from the analyses of the four real datasets, for which it is unfortunately impossible to know the true percentage of outliers because in practice it is impossible to be certain which trials' RTs were contaminated by task-unrelated processes. Therefore, to have a baseline against which to compare these exclusion percentages, I generated four simulated "zero-outlier" (ØO) datasets constructed to be as comparable as possible to the four original datasets and yet known to be free of outliers. First, for each participant in each dataset, separately for the word and nonword conditions, parameters of the best-fitting ex-Gaussian distribution were estimated by maximum likelihood from the observed RTs for that participant and condition. The ex-Gaussian distribution with these parameters was taken to represent the complete distribution of RTs that could occur from task-based processing, and a new set of simulated RTs was randomly generated for each participant and condition from this ex-Gaussian distribution.[7] Thus, all of the generated RTs came from true ex-Gaussian underlying distributions, implying that there were no outliers (i.e., values resulting from task-unrelated processing) in these datasets. Using separate word and nonword parameter estimates for each participant and condition ensures that the simulated datasets would reflect realistic participant-to-participant variation in RT distributions (i.e., mean, variance, and skewness) and in word/nonword effect sizes.

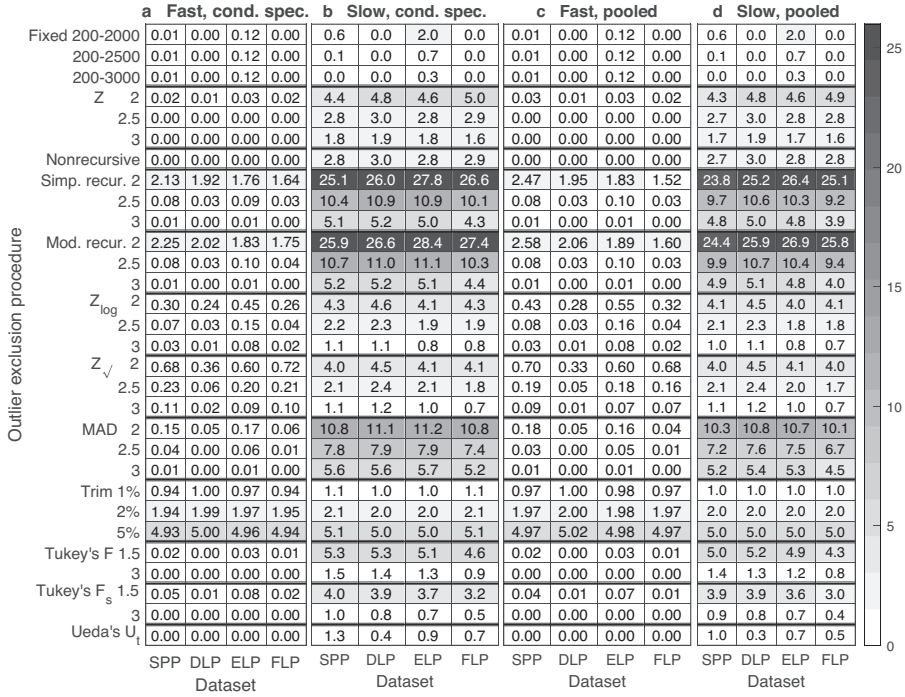
Using each of the four ØO datasets, 5,000 experiments were simulated for analysis using each of the OEPs, and Figure 3 shows the percentages of trials excluded by each OEP from these datasets. These baseline exclusion percentages show that the various OEPs do exclude RTs from true underlying ex-Gaussian distributions—especially as slow outliers—even when all of the RTs are generated from ex-Gaussian task-related processing distributions, with no outlier RTs generated by anomalous processes.

For every OEP, there is a remarkable similarity between the percentage of outliers excluded from the original datasets and the

[7] To mimic the allowable RT range in each dataset, the simulated RTs were generated from a truncated ex-Gaussian distribution, with truncation at the lower limit of 150 ms and at the upper limit corresponding to the maximum RT in each dataset shown in Table 2.

Figure 1

Four Heat Maps (a–d) Showing the Mean Percentages of RTs Excluded as Fast (a and c) and Slow (b and d) Outliers by Each of the Different OEPs for Each of the Different Datasets, Averaging Across the Word and Nonword Conditions



Note. Each OEP was applied separately to the RTs of each participant in each dataset, with either condition-specific application to the word and nonword RTs separately (cond.-spec., a and b) or pooling across word and nonword RTs (c and d). RT = reaction time; OEP = outlier exclusion procedure.

percentage excluded from the ØO datasets constructed to be free of outliers, and this is the case for both condition-specific and pooled application. This suggests that the original datasets did not actually contain many real outliers; if they had, then the OEPs should have identified more outliers in the real datasets than in the ØO datasets constructed from them. But if there were relatively few outliers in the original datasets, then the most accurate OEPs are the ones excluding very small percentages of outliers, and the OEPs excluding larger percentages evidently often misclassified valid RTs as outliers, in some cases excluding a disturbingly large proportion of the trials.

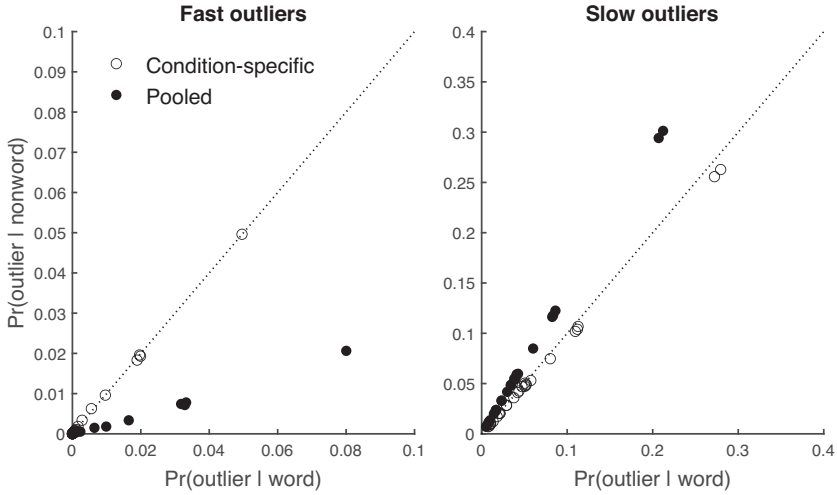
Item Analyses

For each of the letter strings used in these lexical decision task studies, it is possible to compute the mean RT to that string (across participants), excluding the trials classified as outliers so that the outlier RTs themselves would not contaminate these item means. Items that elicit on-average especially fast or slow responses can be assumed to be easier or more difficult than items for which decisions are made at more intermediate speeds. It is also possible to compute, for each letter string, the percentages of trials with that string excluded as fast and slow outliers. The correlation across items of these mean RTs and exclusion probabilities will then reveal whether there is any relationship between an item's difficulty and the probability of its exclusion as an outlier. These correlations across stimulus items were computed separately for word and nonword stimuli within each of the datasets, separately for condition-specific and pooled OEP applications. The two application types produced similar correlations in all cases, and the averages of their correlations are shown in Figure 4. With degrees of freedom based on the number of items across which each correlation was computed, correlations of $|r| > .07$ are statistically reliable for the SPP dataset and those of $|r| > .022$ are statistically reliable for the other three datasets ($p < .01$). Thus, many of the correlations shown in the figure are far too strong to be explained as chance results. For most OEPs, there was a strong positive correlation between the item's mean RT and the proportion of that item's trials that were classified as a slow outlier. These correlations show that inherently relatively difficult items are more likely to be classified as slow outliers, which means that the likelihood of being classified as an outlier depends on item difficulty. The strong associations between item difficulty and probability of outlier classification are extremely problematic for the notion that outliers are generated by task-unrelated processes such as distraction, which should presumably be unrelated to item difficulty. Instead, these correlations provide further evidence that the OEPs sometimes misclassify relatively fast and slow valid RTs as outliers.

One might try to account for the positive correlations of item difficulty and slow outlier classifications by assuming that the slow outlier RTs are the sums of the times needed for task-unrelated processing (e.g., distraction) followed by normal task-related processing

Figure 2

Scattergrams of the Probabilities That Word and Nonword Trials Were Excluded as Fast and Slow Outliers With Condition-Specific and Pooled Application of the Adaptive OEPs



Note. Each point is one OEP, plotted according to its probability of classifying word and nonword trials as outliers. For each OEP, the plotted probabilities were computed separately for each of the four datasets, and the points show the averages of these probabilities across datasets. OEP = outlier exclusion procedure.

(e.g., Ratcliff, 1993). If the times lost to distraction were small, RTs to relatively fast items might fit in with the rest of the observed RT distribution and thus not be classified as outliers, whereas RTs to relatively slow items—after being lengthened further by distraction—would fall outside the OEP's exclusion limits and be classified as outliers. This could explain the positive correlations of item difficulty with slow outlier proportion shown in Figure 4c and 4d. More detailed analysis shows, however, that for many OEPs to fully explain the observed results with this account would require that distraction affected more than 10% of trials, which is an unrealistically large percentage for these lexical decision studies testing normal adult participants under controlled laboratory conditions. Specifically, looking only at the slowest 5% of items, many OEPs classified the RTs to these items as slow outliers in more than 10% of trials (with the maximum across OEPs exceeding 50%). But if task-unrelated distraction slowed responding on more than 10% of trials for the difficult items, it must have done so for all items, which is extremely implausible.

The analogous negative correlations between an item's mean RT and its probability of classification as a *fast* outlier, shown in Figure 4a and 4b, also conflict with the idea that the RTs classified as outliers are invalid values generated by some task-unrelated processes, and they cannot be explained by the same alternative account. With few exceptions, these correlations are much weaker than the correlations with slow outlier classifications in Figure 4c and d, but that is probably because most OEPs classified so few RTs as fast outliers in the first place. For the OEPs that produced more fast outliers (e.g., Trim 5%), the correlations were reasonably strong. It is difficult to see how these correlations could be explained as the combined effects of task-related and -unrelated processing, like the distraction account of the slow outlier correlations, since fast outliers are thought to reflect anticipations and other task-unrelated events that happen before the stimulus is processed.

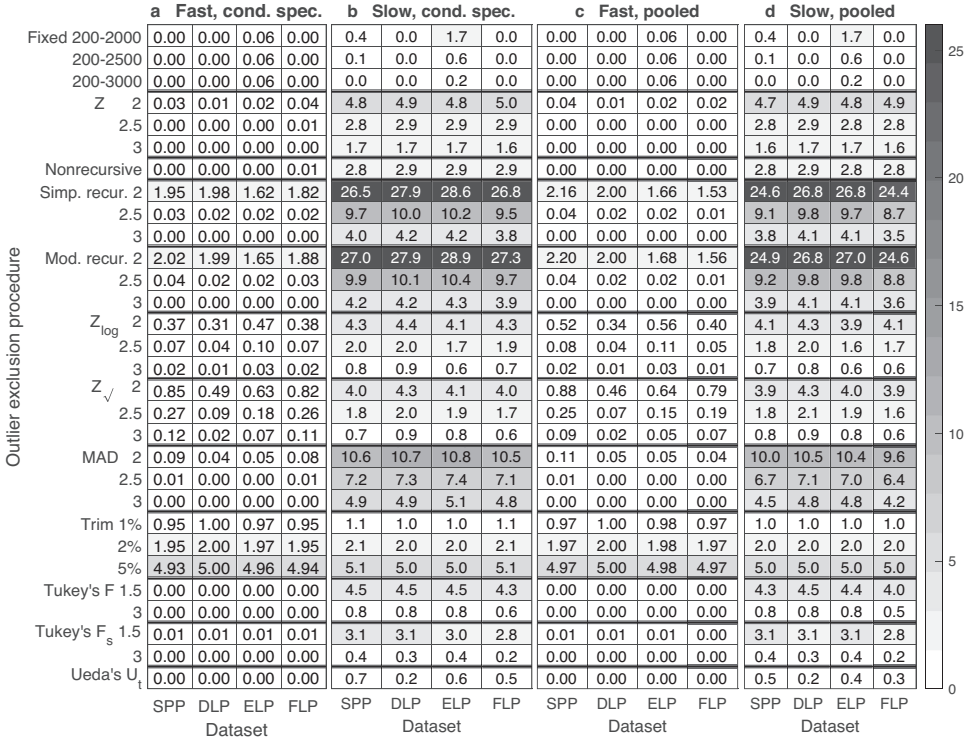
Overall, these item analyses suggest that many of the RTs excluded as fast and slow outliers were actually just fast and slow valid RTs—not actual outliers determined by task-unrelated processes. This is problematic for the OEPs, of course, because all valid RTs should be included in the analysis to obtain the most accurate overall conclusions.

Analyses of Subsamples

Figure 1 shows the percentages of outliers excluded by each OEP from datasets having hundreds of trials per participant in each condition, as is typical in studies with only a few conditions (e.g., words and nonwords). It is also important to examine the performance of the OEPs when used with smaller numbers of trials per condition, however, because such studies are run as well, for many possible reasons. For example, researchers may want to investigate several conditions in a within-subjects design (e.g., a visual search task with 18 conditions defined by Three Types of Targets × Six Stimulus Conditions; Krummenacher et al., 2002). With a maximum of 600–800 trials typically available in a single-session study for practical reasons, it may only be possible to get tens—not hundreds—of trials in each of the many conditions. Another possibility is that the experimental comparison of interest may involve conditions with inherently low numbers of trials; for example, when studying RTs at restricted levels of practice (e.g., Beesley et al., 2016; Mazor & Fleming, 2022) or studying RTs to rare stimuli, responses, or conditions (e.g., Mowrer et al., 1940; Sali et al., 2022). In some designs, the experimental manipulation may require a training or habituation phase, after which the crucial comparison is only possible in a relatively short testing or transfer phase (e.g., Burke & Roodenrys, 2000; Lubczyk et al., 2022), or the testing session may have to be relatively short because the participant population of interest does not have enough patience or stamina for testing hundreds of trials

Figure 3

Four Heat Maps (a–d) Showing the Mean Percentages of RTs Excluded as Fast (a and c) and Slow (b and d) Outliers by Each of the Different OEPs for the Zero-Outlier Datasets Described in the Text, Averaging Across the Word and Nonword Conditions



Note. The format is the same as that in Figure 1.

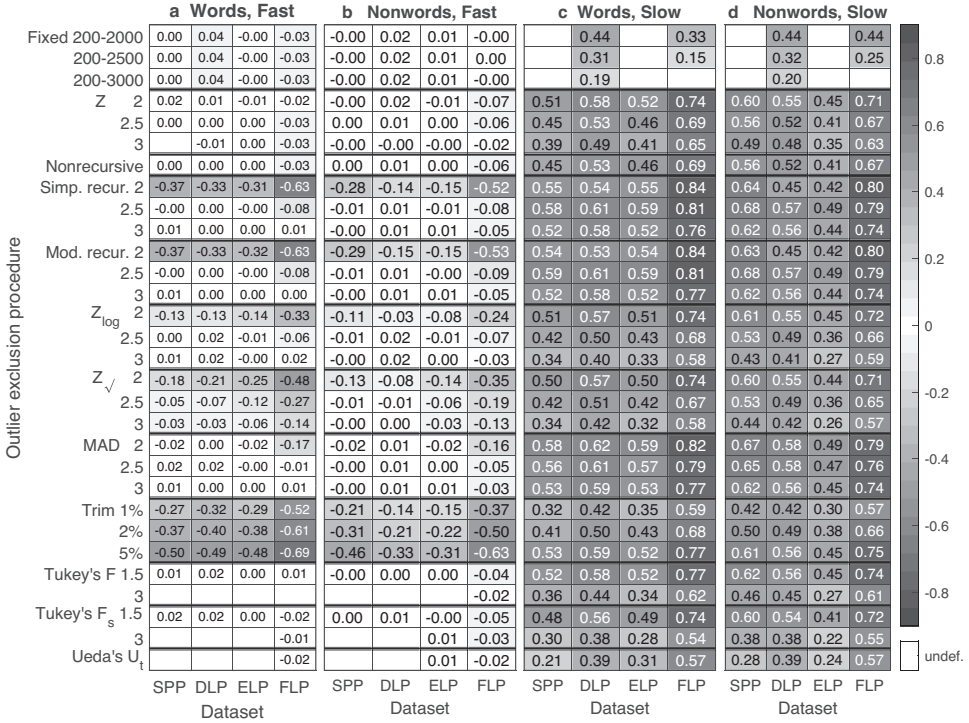
(e.g., children, elderly, patient groups). In all of these cases and more, the number of trials per participant in each condition will be much smaller than the numbers in the SPP, DLP, ELP, and FLP datasets analyzed to this point.

Having a smaller number of trials per condition could substantially influence the performance of the adaptive OEPs, because their outlier classifications of a given RT depend on statistics computed from the other RTs in the sample, which are inherently more variable when based on fewer trials. Using the *Z*-score procedure with a cutoff of 2.5, for example, an RT of 1,000 ms would be excluded from a sample with mean and standard deviation 800 and 100 but not from one with mean and standard deviation 780 and 80. Fluctuations of this magnitude in the relevant sample statistics would be common if there were relatively few trials per condition, so it is important to see how sensitive the performance of each OEP is to random fluctuations in the sample statistics on which it depends. This would be particularly important for condition-specific OEPs, since these base outlier classifications on smaller numbers of trials than pooled ones do.

Figure 5 illustrates the effect of this random sample-to-sample variation on one of the types of scores used to classify each RT as an inlier or outlier (i.e., *Z*-scores). Specifically, for both condition-specific and pooled computations, it shows how the *Z*-scores of two particular RTs would vary randomly across different random samples of trials from a given participant. This figure was constructed starting with the full sample of 817 word-condition RTs from a single participant in the SPP study. Two of the RTs from that sample, 836 and 923 ms, were arbitrarily chosen for examination; these RTs had condition-specific *Z*-scores of 1.87 and 2.52 within the participant's full sample of word RTs and corresponding pooled *Z*-scores of 1.60 and 2.22 considering the participant's nonword RTs as well. For each of these two RTs separately, several sets of 50,000 subsamples of the participant's complete sample of RTs were randomly generated subject to the constraint that the RT being examined (i.e., 836 or 923) belonged to the subsample. In one set of simulations, that RTs' condition-specific *Z*-score was computed within each of these subsamples, and the distribution of condition-specific *Z*-scores was tabulated across the subsamples. The four condition-specific histograms in Figure 5a–5d show the obtained variation in these *Z*-scores for each of the two RTs, separately for subsamples of 20 and 80 trials. Most importantly, the figure illustrates how the classification of a given RT as an inlier or outlier would depend heavily on the other RTs in its sample, even for samples of 80 trials. With the condition-specific *Z* 2 or *Z* 2.5 OEPs, for example, the RT of 836 would be classified as an inlier in the full set of trials, but it would often be classified as an outlier in a smaller sample, because its smaller sample *Z*-score would often exceed the 2.0 or 2.5 cutoff. Conversely, with these OEPs the RT of 923 would be classified as an outlier in the full set of trials, but it would often be classified as an inlier in a smaller sample, where its *Z*-score would often fail to reach

Figure 4

Correlations Across Items Between the Mean of All Participants' RTs to the Item and the Proportion of Participants for Whom That Item Was Excluded as an Outlier, Separately for Each Combination of Fast (a and c) and Slow (b and d) Outliers, Dataset, and Words Versus Nonwords



Note. Empty cells indicate combinations for which a correlation was undefined (undef.) because no trials were excluded as outliers. RT = reaction time.

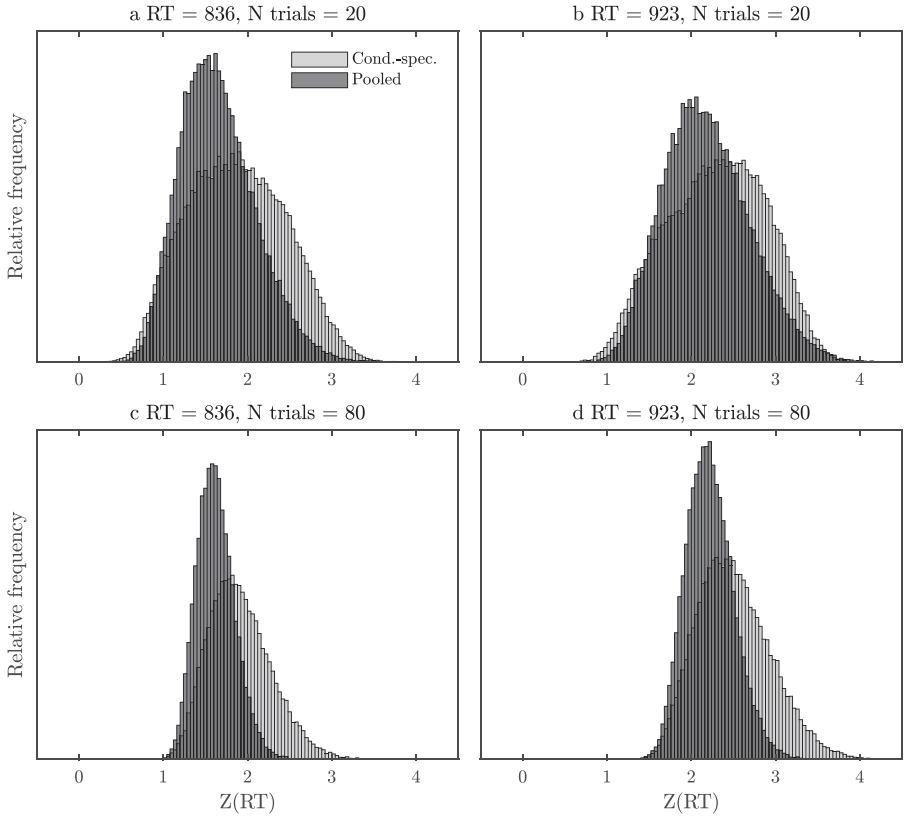
the 2.0 or 2.5 cutoff. The pooled histograms in Figure 5a–5d display the analogous variation with pooled OEP application. These simulated *Z*-scores were also based on samples of 20 or 80 trials per condition except that in these simulations the *Z*-scores were computed relative to the mean and standard deviation of the 40 or 160 sampled trials pooled across both conditions. As predicted, the *Z*-score of a given RT is less variable when computed with the pooled approach than when computed with the condition-specific approach, because the sample mean and standard deviation used in computing *Z* are more stable with the larger pooled set of trials. Even so, the random variation in *Z* with the pooled approach is large enough to be problematic with respect to the identification of outliers. This random variation can of course be decreased by increasing the number of trials, as would also be true for the condition-specific approach. In sum, although these are only examples of *Z*-score variation for two particular RTs from a specific participant, it is easy to see that having a random sample of trials would result in analogous worrisome variation in the individual-trial outlier classifications with all of the adaptive OEPs (i.e., Z_{log}, $Z_{\surd}$, MAD, etc).

Simulations Examining Overall Performance of OEPs

To investigate the overall performance of the OEPs with smaller numbers of trials, simulations were carried out using randomly selected subsamples of the participants and trials from each dataset. Since these were real datasets, these subsamples are equivalent to random selections of data values from each participant's natural RT distribution, as could be obtained in studies with fewer trials per condition if there were no practice effects. The performance of each OEP was examined in $4 \times 3 \times 7 = 84$ different sets of simulation conditions defined by combinations of the four different original datasets, three different sample sizes of participants per simulated experiment (i.e., 10, 20, or 40 participants per simulated experiment, except simulations with the DLP dataset used a maximum of 39 participants—the number available in that dataset), and seven different numbers of trials (10, 20, 40, 60, 80, 120, or 160) per participant in each of the two experimental conditions (i.e., word vs. nonword). In total, 10,000 replications of an experiment were simulated under each of these 84 conditions for each of the 58 OEPs. This section reports tabulations of these simulation results examining how many and which trials were excluded as outliers. These tabulations were pooled across experiments with different numbers of participants, because the OEPs are applied to the trials of each simulated participant separately, regardless of how many other participants re included in the study. In contrast, the following section reports results concerning statistical power, and in this case, the results of simulations with different numbers of participants were examined separately because the number of participants affects power.

Figure 5

Illustrations of the Variation in Condition-Specific (Cond.-Spec.) and Pooled Z-Scores for the RTs of 836 (a and c) and 923 ms (b and d) Across Random Subsamples of 20 (a and b) or 80 (c and d) Trials Per Word/Nonword Condition From a Single Participant's Full Sample of RTs



Note. RT = reaction time.

The basic procedure was the same for each of the 10,000 simulated experiments under each of the 84 simulation conditions. First, depending on the condition, either 10, 20, or 40 participants were selected from a given dataset, and then subsamples of either 10, 20, 40, 60, 80, 120, or 160 trials per word/nonword condition were selected from the actual observed word and nonword RTs of each of those participants.[8] Next, the OEP being examined was applied in either a condition-specific or pooled manner to the word and nonword trials of each participant to see which trials would be classified as outliers, just as in the analyses of the complete datasets presented earlier (and just as they would be applied in actual experiments with these smaller numbers of trials per participant). After the exclusions had been determined, the mean RTs of the remaining trials were computed for each participant and condition. Then, to investigate power, the word/nonword effect was evaluated in each simulated experiment using a paired *t*-test computed using the participants' mean RTs—after exclusion—in the word and nonword conditions. A real effect is of course known to be present in each of the original datasets from which the participants and trials are sampled, though the effect size differs across studies (see Table 2).

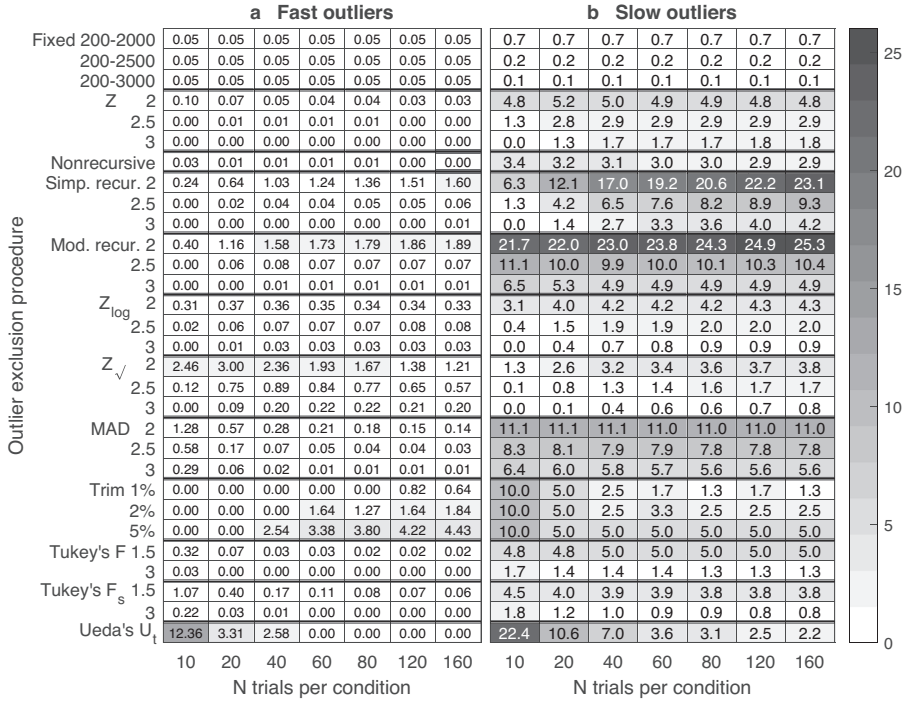
Percentages of Outliers Excluded

Figure 6a and 6b shows the percentages of RTs excluded as fast and slow outliers, respectively, for each combination of OEP and the number of trials, with condition-specific application of the OEPs. Naturally, with the largest N of 160 trials, the exclusion percentages are similar to those found with the full datasets (Figure 1a and b), since the exclusions were determined with mostly the same trials in both of those cases. For smaller numbers of trials, however, the exclusion percentages change substantially for some OEPs—in some cases increasing as N decreases (e.g., Ueda's U_t), in some cases decreasing as N decreases (e.g., simple recursive 2), and in some cases increasing for fast outliers but decreasing for slow outliers (e.g., $Z_{\sqrt{}}2$). Changes in OEP behavior with N are problematic for both within-experiment and between-experiment comparisons of conditions with different numbers of trials (e.g., Miller, 1991), so OEPs that are insensitive to N would be preferable. The exclusion procedures with fixed cutoffs (e.g., 200–2,000) are clearly the best by this criterion, since they exclude exactly the same RTs regardless of how many or which other RTs are in the sample, and the fixed-cutoff percentages shown in Figure 6 reflect the averages across datasets of the corresponding percentages in Figure 1a and 1b.

[8] Both participants and trials were randomly selected without replacement in most simulations. The exceptions were the simulations with the DLP dataset and 39 participants; since this was the full set of participants, participants were selected with replacement in these simulations.

Figure 6

Percentages of RTs Classified as Fast (a) and Slow (b) Outliers in Randomly Selected Subsamples of N Trials Per Participant Per Condition, as a Function of the Outlier Exclusion Procedure (Condition-Specific Application) and N



Note. Percentages were averaged across the four original datasets and the word/nonword conditions. RT = reaction time.

The analogous percentages excluded with pooled OEP application are not shown because they were very similar to those with condition-specific application (for fast outliers, a mean square difference between condition-specific and pooled values of 0.8% and a rank order correlation of 0.85; for slow outliers, corresponding values of 2.2% and 0.91). As predicted on theoretical grounds, however, there were substantial differences between condition-specific and pooled application with respect to the exclusion of RTs from the fast word versus slow nonword trials (not shown), analogous to those shown in Figure 2. Regardless of the number of trials per condition in the subsample, the percentages of word- and nonword-trial RTs classified as outliers were always similar to one another for each OEP with condition-specific application, but with pooled application, the OEPs always tended to exclude more word RTs as fast outliers and more nonword RTs as slow outliers.

Stability of Outlier Classification Across Sample Sizes

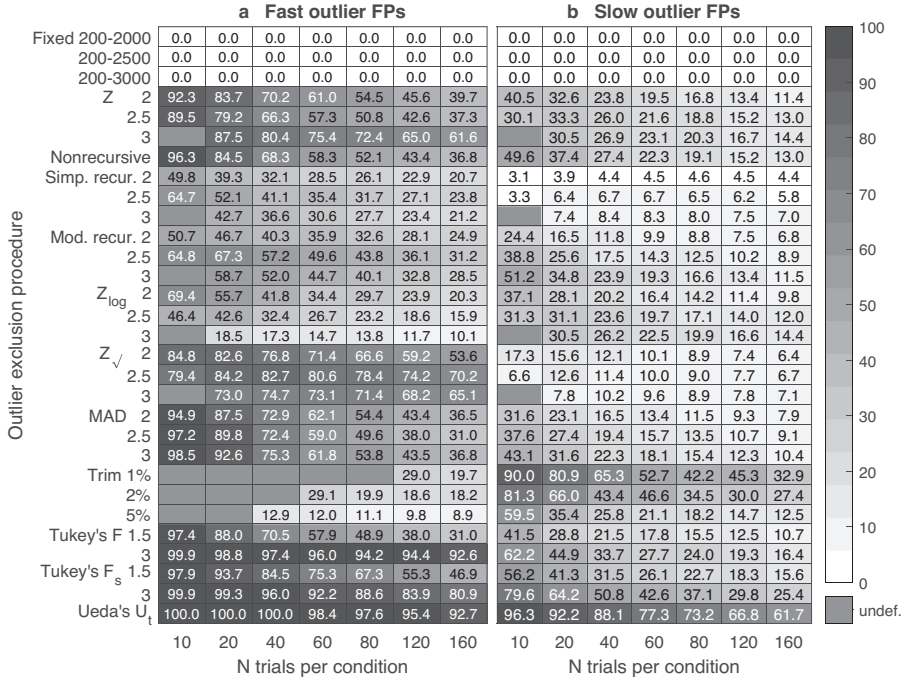
With real datasets, it is impossible to be certain about each OEP's *accuracy* in classifying outliers, because the true inlier/outlier status of each RT is unknown. It is possible, however, to check the *stability* of the OEP's classifications across samples with different numbers of trials. In theory, a given RT from a particular participant in a particular condition is either an inlier or an outlier, so an ideal OEP would classify it the same way regardless of the other RTs in the sample. This is guaranteed to be the case for the OEPs using fixed RT cutoffs (e.g., 200–2,000), but it cannot be perfectly true for the adaptive procedures due to sampling variability in the statistics used to compute their RT cutoffs (e.g., Figure 5).

To examine the stability of each OEP's inlier/outlier classifications, the classification of each RT in the full set of all of the participant's RTs within a condition was designated as the "best guess" outlier classification of that RT for each OEP, and the classifications of RTs as inliers versus outliers within the subsamples were then checked to see how well these agreed with the best guess classifications.

Figure 7 shows the percentages of subsample outlier classifications that were false positives (FPs) using condition-specific application of the adaptive OEPs—that is, these are the percentages of the RTs classified as outliers in the subsamples that had been accepted as inliers in the best-guess classifications using the same OEP but considering all of the trials from the same participant and condition. In essence, these FP percentages can be interpreted as the percentages of all outliers excluded from the subsamples that should actually have been retained in the analysis. Unfortunately, it is clear that almost all of the adaptive OEPs have disturbingly large FP rates for fast outliers, for slow outliers, or for both, especially in samples with fewer trials. In the worst cases, more than 90% of the trials discarded from the subsamples as outliers should actually have been accepted as inliers judging by the best guesses from the full set of trials. The problem is of course most serious when there are relatively few trials in the subsample,

Figure 7

Percentages of the RTs Classified as Fast (a) and Slow (b) Outliers in the Subsamples of RTs That Were FPs (i.e., Had Been Classified as Inliers in the Full Set of RTs), as a Function of the Outlier Exclusion Procedure (OEP; Condition-Specific Application) and the Number of Trials in the Subsample



Note. Percentages are averaged across the four original datasets and the word/nonword conditions. The empty cells reflect cases where these percentages could not be computed because no trials were classified as outliers in the subsamples. See Figure 9a and 9b for comparisons with pooled-application OEPs. RT = reaction time; OEP = outlier exclusion procedure; FPs = false positives.

but it is remarkable how unstable the results can be even with 160 trials per subsample, at least in the classification of fast outliers. The bottom line is that with smaller samples almost all of the adaptive OEPs throw away many trials as outliers that they would actually have retained for the analysis if a much larger sample of trials had been available.

Figure 8 shows the percentages of RTs classified as outliers in the best-guess analyses that were missed—that is, classified as inliers—in the analyses of subsamples of trials from the same participant and condition, also using condition-specific application of the OEPs. Again, except for the fixed-cutoff OEPs which necessarily exclude the same RTs regardless of the sample being analyzed, it is clear that there is poor agreement between full-sample and subsample outlier classifications. More specifically, few of the RTs classified as outliers in the full samples were also classified that way in the subsamples. Even with samples of 160 trials, the miss rates are greater than 10%—which seems large enough to be concerning—for most OEPs. The problem is naturally more severe with smaller subsamples, even approaching 100% misses with 10 trials for several OEPs. Evidently, with a small sample of trials an extreme RT can have a large enough influence on an adaptive OEP's sample statistics (e.g., sample mean and standard deviation for *Z*-score-based OEPs) to protect itself from exclusion. The modified recursive procedure of Van Selst and Jolicœur (1994) that was designed to ameliorate this number-of-trials problem for the slow outliers is partially successful but is far from escaping the problem entirely. Thus, as was the conclusion in the analysis of FPs, OEPs are not stable across sample sizes with respect to the classification of particular RTs as inliers versus outliers.

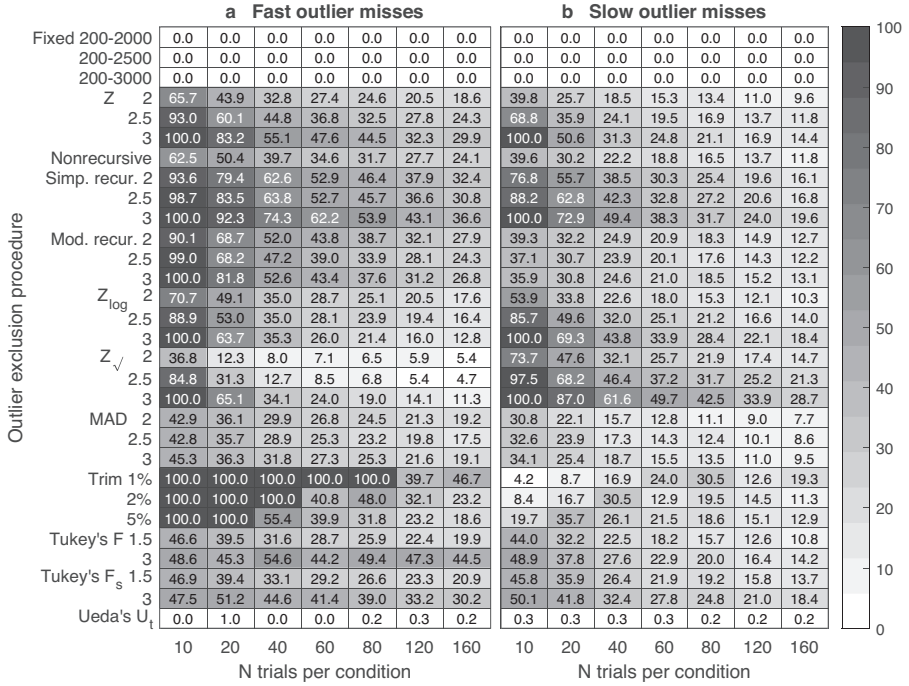
Figure 9 shows a comparison of the condition-specific and pooled approaches with respect to their respective rates of FPs and misses. Each point in the figure represents a particular combination of one of the 27 adaptive OEPs and one size of RT subsample (i.e., 10, 20, 40, 60, 80, 120, or 160 trials per condition), with the FP or miss probabilities for the condition-specific and pooled versions of the OEP shown on the horizontal and vertical axes, respectively. Most of the points in these scattergrams lie below the diagonal line of equality, indicating that there were fewer FPs and misses with pooled than with condition-specific OEP application, presumably because of the greater stability of sample statistics with the pooled approach. There were still high percentages of both types of incorrect classifications with the pooled OEPs, however, so neither approach performs well in absolute terms.

Which RTs Are Excluded?

To get further insight into the performance of the different OEPs and the causes of the disagreements between the full-sample and

Figure 8

Percentages of the RTs Classified as Fast (a) and Slow (b) Outliers in the Analysis of All RTs That Were Missed (i.e., Classified as Inliers) in the Analyses of Subsamples, as a Function of the Outlier Exclusion Procedure (OEP; Condition-Specific Application) and the Number of Trials in the Subsample



Note. Percentages are averaged across the four original datasets and the word/nonword conditions. See Figure 9c and 9d for comparisons with Pooled-Application OEPs. RT = reaction time; OEP = outlier exclusion procedure.

subsample outlier classifications, it is illuminating to look more closely at which particular RT values were excluded as outliers by each of the OEPs with each of the different sample sizes. For example, if 2% of the RTs at each end of an RT distribution are outliers, then ideally an OEP should exclude RTs with percentile ranks in the ranges of 0–2 and 98–100, as measured relative to the full set of RTs from a participant (pooled) or participant and condition (condition-specific). The preceding FP and miss results show that the OEPs are not doing that perfectly, but they do not indicate how far off-target the OEPs actually are. For example, might an RT with a percentile rank of 85 occasionally be excluded as a slow outlier in a small sample of RTs if there were actually only 2% slow outliers? This would clearly be incorrect, and it would cause the mean RT to be underestimated. Similarly, might an RT with a percentile rank of 99.9 occasionally be accepted? This would also be incorrect and would cause the mean RT to be overestimated. This section addresses these questions by examining more closely the likelihood of accepting or excluding RTs at different percentile ranks within the overall distribution to get a sense of the most serious errors that each OEP might make.

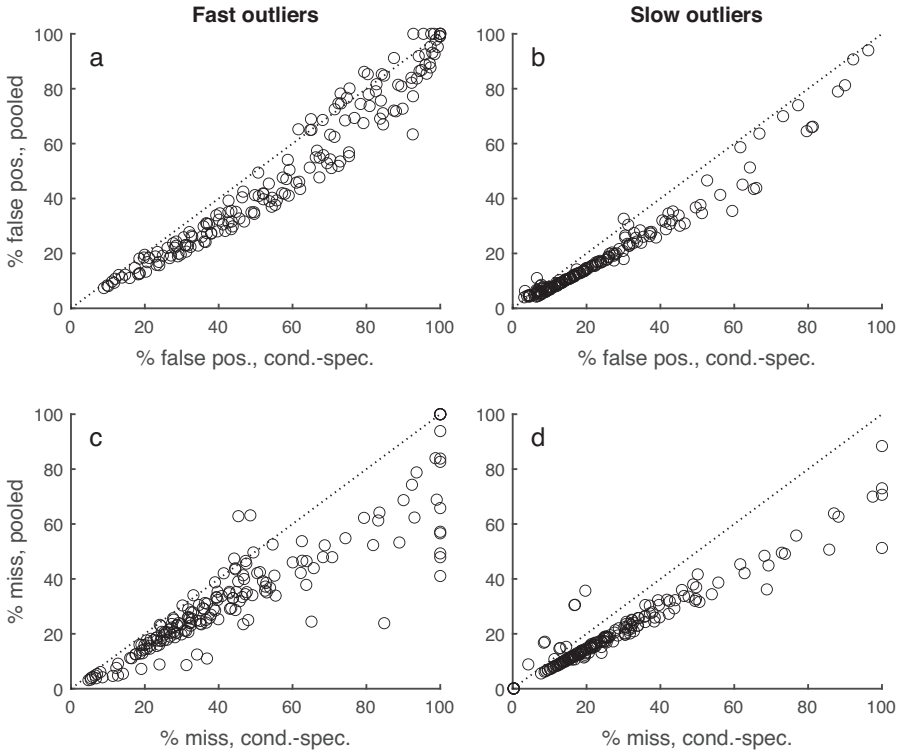
Figure 10 illustrates one way of doing this using the slow outlier classifications for the *Z* 2.5 and simple recursive 2.5 OEPs as examples,[9] and the results with condition-specific and pooled application are very similar. Each RT from each participant was assigned to a percentile rank bin (0–1, 1–2, …, 99–100) based on its percentile rank within the set of all RTs from that participant and condition for the condition-specific OEPs or within the set of all RTs from that participant across both conditions for the pooled OEPs. This percentile-rank binning provides a convenient method of normalizing the RT values across participants, datasets, and conditions (Miller, 2021). The vertical axis in Figure 10 shows, for each percentile rank bin, the proportion of that bin's RTs that were excluded as slow outliers by the OEP in analyses with the indicated number of trials (i.e., all trials vs. subsamples of 20 or 80 trials), averaging the proportions for each bin across datasets and word/nonword conditions.

First, Figure 10 illustrates why the *Z* 2.5 OEP excludes many fewer slow outliers in the analyses of all trials than the simple recursive 2.5 OEP (cf., Figure 1). When considering all trials, *Z* 2.5 hardly ever excludes any RTs below the 95th percentile of the participant's RT distribution, whereas simple recursive 2.5 sometimes excludes RTs all the way down to the 85th percentile and even beyond. Implicitly, then, the *Z* 2.5 OEP suggests that at most 5% of the trials were outliers for any participant, whereas the simple recursive 2.5 OEP suggests that as many as 15% of the trials were slow outliers for some participants. The recursive procedure tends to exclude

[9] Analogous figures for the other OEPs are provided in the online supplemental materials.

Figure 9

Scattergrams of the Percentages of Fast and Slow Outlier Classifications That Were False-Positives (False Pos.; a, b) and Misses (c, d) With Condition-Specific (Cond.-Spec.) and Pooled OEP Application in Subsamples of RTs



Note. Each point is one combination of OEP and number of trials, plotted according to the corresponding average percentages of FPs or misses, averaging across the four datasets, and word/nonword. RT = reaction time; OEP = outlier exclusion procedure; FPs = false positives.

RTs down into smaller percentiles because the exclusion of any large RT in the first pass of the procedure reduces the standard deviation of the remaining RTs. This tends to increase the *Z*-scores of the large remaining RTs in the second and later passes, making it more likely that one of them will be excluded at the next step.

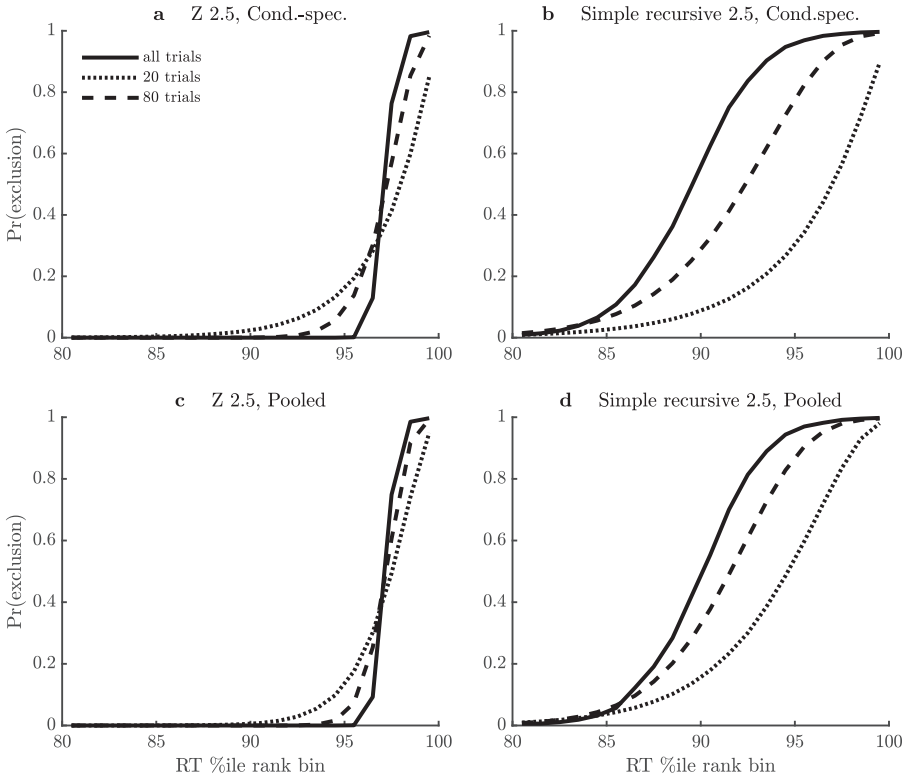
Second, Figure 10 illustrates the different influences of the number of trials on the *Z* 2.5 versus simple recursive 2.5 OEPs. For the *Z* 2.5 OEP, reducing the number of trials essentially smears the boundary between inliers and outliers. This boundary is quite sharp at approximately the 96th–97th percentile in the analysis of all trials, with RTs below that boundary virtually never excluded and RTs above that boundary virtually always excluded. In contrast, with subsamples of 20 trials, RTs as low as the 90th percentile are sometimes excluded, and these are the FPs shown in Figure 7. Similarly, those at the 98th–99th percentiles are sometimes accepted; these are the misses shown in Figure 8. With the simple recursive 2.5 OEP, on the other hand, reducing the number of trials simply reduces the number of slow outliers excluded (cf., Figure 6), causing a lot of misses without increasing FPs.

One way to summarize all of the OEP's curves analogous to those shown in Figure 10 is to find the percentile ranks of the *least* extreme RTs ever classified as fast and slow outliers within subsamples of RTs. With the condition-specific *Z* 2.5 OEP, for example, the curve for 80 trials shows that RTs down to approximately the 90th percentile of the individual participants' RT distributions are occasionally rejected as high outliers; with 20 trials, RTs down even below the 85th percentile are occasionally rejected. If outliers are actually rare in these datasets as expected and as suggested by the evidence presented earlier, then only RTs with the most extreme percentile ranks (e.g., 0–2, 98–100) should actually be classified outliers. Thus, the performance of each OEP can be checked to see whether it in fact excludes only the most extreme RTs—that is, excludes only the fastest RTs as fast outliers and only the slowest as slow outliers.

Figure 11 shows summaries of the percentile ranks of the least-extreme outliers for all of the condition-specific OEPs and sample sizes, averaging across participants, datasets, and word/nonword conditions. To compute these values, the first step was to find—for each participant and word/nonword condition separately—the largest RT ever excluded from a subsample as a fast outlier and the smallest RT ever excluded as a slow outlier. The percentile ranks of these least-extreme RTs within the participant's full set of RTs from that condition (condition-specific) were then recorded, and these least-extreme percentile ranks were averaged across participants, datasets, and word/nonword conditions, separately for the largest fast outliers and the smallest slow outliers. For example, the 15 in the upper left cell of Figure 11a indicates that the condition-specific *Z* 2 OEP excluded RTs as fast outliers up to the

Figure 10

Percentage of Trials Within a Given RT Percentile (%ile) Rank Bin Excluded as Slow Outliers Using Condition-Specific (Cond.-Spec.; a, b) and Pooled (c, d) Application of the Z 2.5 (a, c) and Simple Recursive 2.5 (b, d) OEPs With Analyses Based on All Trials or on Subsamples of 20 or 80 Trials



Note. RT = reaction time; OEP = outlier exclusion procedure.

individual participants' 15th percentiles, on average, when used with subsamples of 10 trials. Since this is an average across participants, though, the largest values are naturally even larger; for some participants, this OEP occasionally rejected an RT *greater than the participant's overall median* as a fast outlier when that RT was included in a certain subsample of 10 trials.

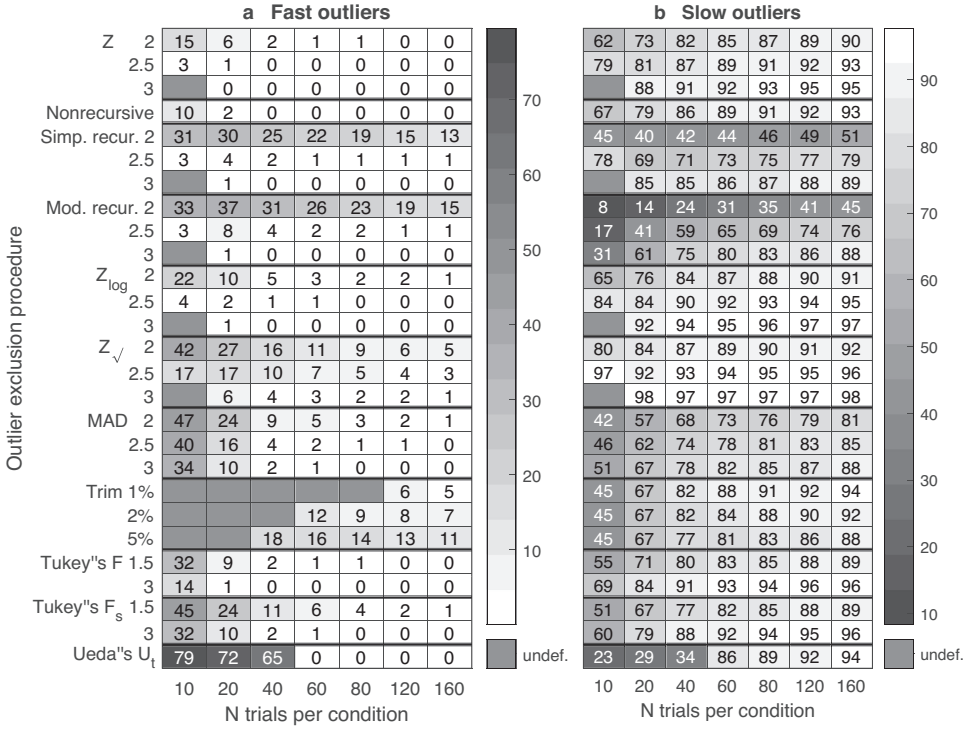
The performance of many condition-specific OEPs is dramatically worse with respect to the exclusion of slow outliers, as is shown in Figure 11b. If there is only a small percentage of true slow outliers, as the previous analyses suggest, then the OEPs should only reject RTs at the highest percentiles (e.g., 98–100). Few of the adaptive OEPs come close to satisfying this condition, however. For some of them, FP rejections can occur down to much lower percentiles; even RTs faster than individual's median for a condition can sometimes be rejected as slow outliers in random samples of the participants' RTs. Despite the intuitive appeal of these adaptive procedures, then, it is clear that the inherent sampling variability influencing the rejection cutoffs can lead to the rejection of RTs that cannot possibly be real outliers in the intended sense of that term.

The analysis of least-extreme outliers reveals another small difference between the condition-specific and pooled approaches that can be seen in Figure 12. As in Figure 9, each point represents a combination of one adaptive OEP and one value of the number of trials, and the point is plotted according to the average of the least-extreme outlier percentile ranks computed with each type of application. Pooled application seems better than condition-specific application, because the RTs excluded as outliers with pooled application are a bit more extreme than those excluded as outliers with condition-specific application (i.e., lower percentile ranks for fast outliers and higher ones for slow outliers), presumably due to the more stable sample statistics of the pooled approach. Unfortunately, the improvements obtained with the pooled application are generally small, so pooled application is no panacea for the problem that adaptive OEPs often reject nonextreme RTs that simply look extreme within a particular small set of trials.

Figure 13 shows another summary measure of the sampling-based inconsistency of each adaptive OEP's outlier classifications—a measure that could be called the "spread of inconsistency" (SOI). This measure is based on the fact that relatively more extreme RTs are sometimes accepted as inliers and relatively less extreme RTs are sometimes rejected as outliers. For example, the curve for 20 trials in Figure 10a shows that RTs at the 90th percentile are sometimes excluded—Pr(exclusion) > 0, whereas RTs at the 99th percentile are sometimes accepted—Pr(exclusion) < 1. This OEP's classifications of RTs as slow outliers can thus be regarded as inconsistent across approximately the 90–99 range of percentile bins.

Figure 11

Mean Percentile Ranks of the Largest RTs Excluded as Fast Outliers (a) and the Smallest RTs Excluded as Slow Outliers (b), as a Function of the Number of Trials in a Subsample and the Adaptive Outlier Exclusion Procedure, With Condition-Specific Application



Note. RT = reaction time.

For each sample size and OEP, slow and analogous fast outlier SOI widths were computed separately with each participant's RTs in each condition. The SOI for a slow outlier classification was defined as the range from the smallest RT ever excluded as a slow outlier (across all subsamples with the given number of trials) to the largest RT sometimes accepted as an inlier (across all subsamples with that number of trials) for that participant. For example, if an OEP excluded a certain participant's word condition RT of 1,100 ms as a slow outlier in some subsamples yet accepted a word RT of 1,300 ms as an inlier in other subsamples, then the OEP's slow outlier classifications for that participant were inconsistent over an SOI with a width of 1,300–1,100 = 200 ms. For an OEP with no inconsistency, all RTs rejected as slow outliers would be larger than all RTs accepted as inliers. In this case, the SOI width would be negative or zero, as is necessarily the case for the fixed-cutoff OEPs.

Figure 13 shows the averages of the SOI widths for each adaptive OEP with condition-specific application, averaging across participants, datasets, and conditions; the fixed-cutoff OEPs have been omitted because their widths must be negative or zero. The SOIs are clearly wider with smaller numbers of trials, and this is to be expected because smaller samples of trials have more random variability in the statistics used to determine outlier cutoffs. The SOIs for slow outlier classifications are also much wider than those for fast outlier classifications, presumably as a result of the skew at the high ends of the RT distributions.

Comparing the different OEPs, with large numbers of trials the MAD procedures seem to have slightly narrower SOIs than the others, at least for slow outliers, and trimming seems to have especially wide ones. More broadly, though, it is reasonable to ask whether the absolute magnitudes of all of the SOIs in Figure 13 are just unacceptably large. For example, would one trust an OEP which might reject a 1,000 ms RT as a slow outlier in one sample but accept a 1,750 ms RT as an inlier in another sample from the identical RT distribution? The SOIs of greater than 750 ms in the figure shows that this would happen for many OEPs applied to samples of 10 or 20 trials.

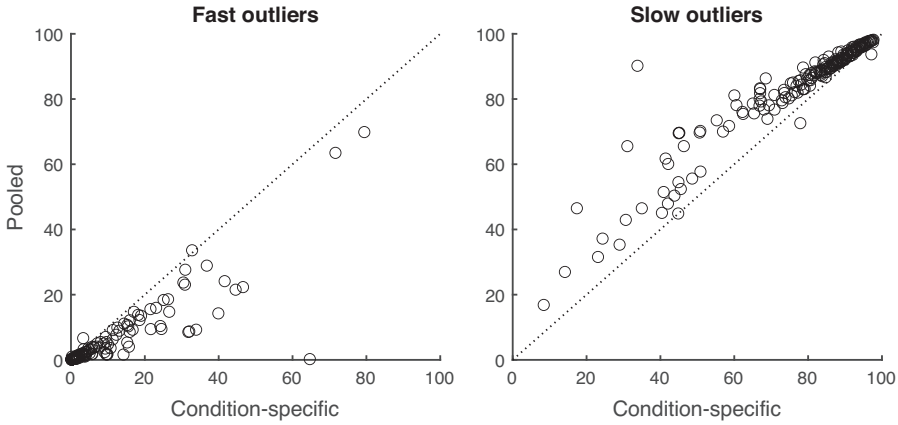
Figure 14 shows scattergrams relating the SOI widths for each adaptive OEP with condition-specific versus pooled application, averaging across participants, datasets, and word/nonword conditions. As expected because of its more stable sample statistics, pooled application works slightly better with respect to this measure (i.e., narrower SOIs as indicated by points below the diagonals), yet its overall SOI widths are still disturbingly large.

Testing for Experimental Effects

Given the erratic, questionable performance of many OEPs in classifying RTs of outliers, it is essential to ask what impact the choice of OEP has when testing for the existence and size of experimental effects. Ideally, researchers would like to use an OEP that would (a) maximize power while keeping Type 1 error rate within the nominal α level (typically 5%), and (b) yield unbiased and low-variance

Figure 12

Scattergrams of the Mean Percentile Ranks of the Largest RTs Excluded as Fast Outliers and the Smallest RTs Excluded as Slow Outliers With Outliers Computed Using Condition-Specific Versus Pooled OEP Application



Note. Each point depicts the result for one adaptive OEP with one value of the number of trials per condition. RT = reaction time; OEP = outlier exclusion procedure.

estimates of the true effect size. It is easy to imagine that many of the OEPs actually work against those aims by discarding potentially large numbers of valid RTs prior to the main analyses.

To address these questions, I investigated the performance of the various OEPs using the same simulated experiments described previously (i.e., simulated experiments with 10–160 randomly selected real RTs per condition for each of 10, 20, or 40 randomly selected participants). To examine power and effect size estimation in within-subjects designs, the trials for the two conditions were randomly selected from the participants' word and nonword RTs, respectively, with power estimated for each combination of numbers of participants and trials by tabulating the proportion of significant word/nonword effects, as described below.[10]

Transposed Datasets

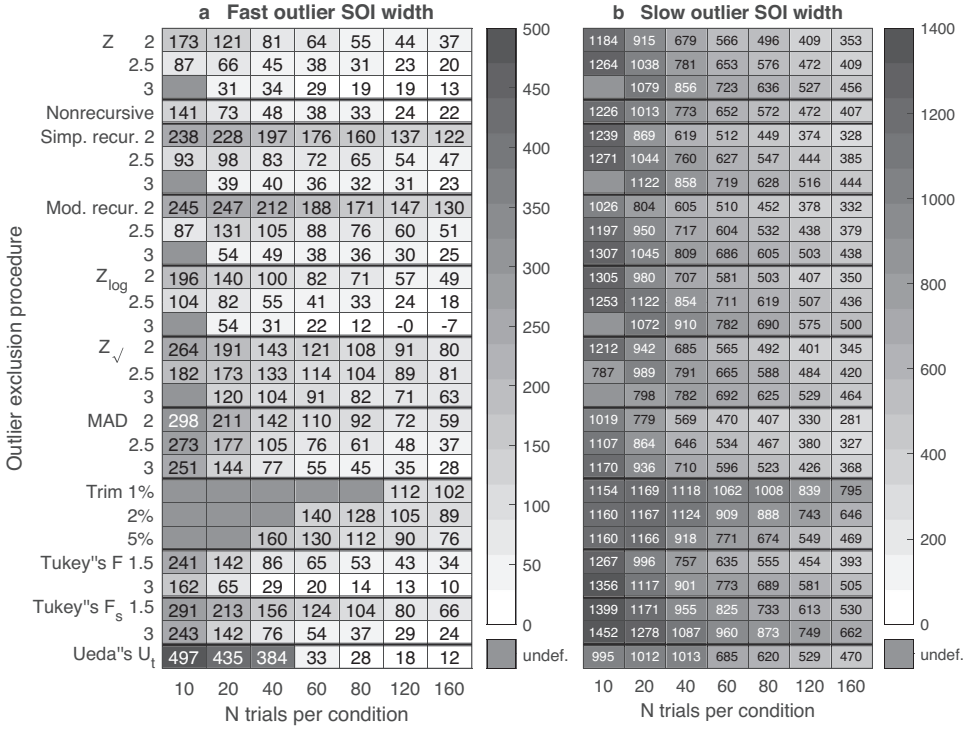
Unfortunately, detailed analyses of the RT distributions in the four original datasets used in this article revealed a serious limitation with respect to the present study's overall goal of seeing how well the different OEPs would perform in detecting experimental effects. Specifically, in all four of the original datasets, the word/nonword effect mainly involved a simple additive increase or "shift" of the RT distribution in the slower nonword condition relative to the faster word condition, with little or no change in skewness. Often, experimental manipulations that increase the mean RT within a condition also increase its variability within that condition (Luce, 1986), and some experimental manipulations increase mean RT *mostly* by "stretching" the RT distribution—that is, by having a small effect on the fastest responses but a large effect on the slowest ones, essentially increasing skewness in the slower condition (e.g., Heathcote et al., 1991; Hockley, 1984; Hockley & Corballis, 1982; Hoedemaker & Gordon, 2017; Low et al., 2002; Moutsopoulou & Waszak, 2012; Palmer et al., 2011; Possamaï, 1991; Singh et al., 2018; Steinhauser & Hübner, 2009). To have a more thorough comparison of OEP performance in detecting experimental effects, it would thus be helpful to have additional datasets in which the experimental manipulation had its effect primarily by stretching rather than shifting the RT distribution in the slower condition. In particular, if the experimental conditions differ in skewness, the power to detect a statistically significant difference may well vary across OEPs because the classification of outliers can be affected by the extent of RT skewness (e.g., Berger & Kiefer, 2021; Miller, 1991).

Since I could not find large publicly available RT datasets with stretch rather than shift effects, I constructed some—one for each of the four datasets shown in Table 2. This was done by transposing the word/nonword effect from a shift effect to a stretch effect, separately for each participant in each dataset, within an ex-Gaussian approximation of each observed RT distribution (e.g., Heathcote et al., 1991). The computational procedure used to do this is described in Appendix A, but the basic idea was to replace each nonword RT for a given participant with a corresponding RT from a "transposed" nonword distribution designed to be as similar as possible to the participant's original nonword RT distribution except with a stretch effect (i.e., on the τ parameter of its estimated ex-Gaussian) instead of the shift effect (i.e., on the ex-Gaussian μ parameter) found in the original dataset. The procedure was carried out separately for each participant in each dataset, and only the participants' RTs in the nonword condition were changed for the transposed datasets relative to the original ones; the word RTs were not altered. The sets of simulations examining power and Type 1 error rate with the transposed datasets were exactly parallel to those carried out with the original datasets, again with 10,000 simulated experiments for each combination of dataset, OEP, number of participants, and number of trials per condition.

[10] A complication arose when tabulating power for some of the OEPs applied in a pooled manner. For some of these, there were occasional simulated experiments in which all of one participant's RTs in a given condition were excluded as outliers, making it impossible to conduct the planned paired *t*-test. These simulated experiments were excluded from the analysis of power.

Figure 13

Mean Widths, in Milliseconds, of the SOIs for Fast (a) and Slow (b) Outlier Classifications as a Function of the Number of Trials in a Subsample and the Adaptive Outlier Exclusion Procedure Applied in a Condition-Specific Manner



Note. SOIs = spreads of inconsistency.

Results

Figure 15 shows the most important results of the simulations examining the performance of the different OEPs in the hypothesis testing situation, summarizing the performance of each OEP with respect to its statistical power, its bias in effect-size estimation, the standard error of its effect-size estimate, and its Type 1 error rate. Each of these aspects of OEP performance will be considered in turn.

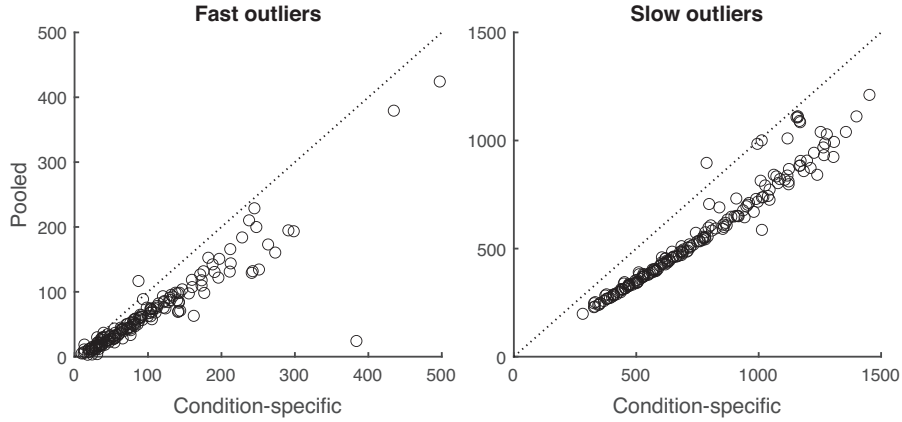
Power

The power of each OEP was assessed as the proportion of the 10,000 simulated experiments in which the participants' mean RTs—after exclusions, if any—yielded a statistically significant paired *t*-test for the word/nonword comparison. If statistical significance had been judged using the traditional significance cutoff of $\alpha = .05$, however, the power values would have been near the ceiling of 1.0 in many of the present simulation conditions (e.g., with 40 participants and 160 trials per condition) due to the large word/nonword differences in the original datasets. To avoid ceiling effects that could obscure the differences among OEPs and to normalize the power computations in a manner that would facilitate comparing the power levels of the different OEPs across the different combinations of datasets, numbers of participants, and numbers of trials, a separate specially-selected α level was chosen for use with each combination instead of the standard $\alpha = .05$.

Specifically, the α for a given combination was chosen to be the median *p* value obtained across the simulated experiments with that combination when analyzed with no exclusions. For example, 10,000 simulated experiments with 20 participants from the original SPP dataset and 20 trials per participant in each condition were analyzed with no exclusions. The 10,000 *t*-tests from these analyses yielded 10,000 *p* values, and the median of these no-exclusion *p* values was .0010784. The value .0010784 was thus chosen as the critical α level for all simulated SPP experiments with 20 participants and 20 trials per participant in each condition. This meant that the power with no exclusions was necessarily 0.5, because exactly half of its experiments produced $p < .0010784$, allowing the null hypothesis (HØ) to be rejected at this critical α. The same critical $\alpha = .0010784$ level was then used for all of the other OEPs in experiments simulated with 20 SPP participants and 20 trials per participant. OEPs that improved power relative to no-exclusions for that combination would produce more experiments with $p < .0010784$, thus rejecting HØ more than half the time and thus yielding power values greater than 0.5, whereas those with worse power would produce fewer $p < .0010784$ values and thus reject HØ less often at this critical α, resulting in power values less than 0.5. In essence, then, the power of each OEP was estimated as the proportion of times that the analysis with that OEP yielded a *p*-value less than the median *p*-value obtained under the same conditions with no exclusions. Most importantly for the present purposes, power levels assessed in this manner are free of floor and ceiling effects and are

Figure 14

Scattergrams of the Mean Widths, in Milliseconds, of the SOIs Regarding Fast and Slow Outlier Exclusions With Outliers Computed Using Condition-Specific Versus Pooled OEP Application



Note. Each point depicts the result for one adaptive OEP with one value of the number of trials per condition. SOIs = spreads of inconsistency; OEP = outlier exclusion procedure.

appropriate for comparing the different OEPs, because the same critical α level was used for all OEPs in any given simulation condition. Note that the power levels assessed in this fashion are also comparable across sample sizes and datasets, because they have all been normalized so as to produce power 0.5 with no exclusions in all situations. A table showing the critical α level chosen for each combination is shown in Appendix B. Not surprisingly, the critical α levels needed to produce $\Pr(p < \alpha) = 0.5$ with no exclusions vary over multiple orders of magnitude from the combinations with the smallest effects and fewest participants and trials (e.g., DLP dataset with 10 participants and 10 trials) to the combinations with the largest effects and most participants and trials (e.g., SPP dataset with 40 participants and 160 trials).

Figure 15a shows the average of the power values computed in this fashion, averaging across different datasets and different numbers of participants and trials. When they are applied to the original datasets in a condition-specific manner, the OEPs generally provide less power than using no exclusions at all (i.e., they reject HØ less than half of the time). This is disappointing, because OEPs are widely used by RT researchers to remove noise and thus provide more powerful tests for experimental effects. Fortunately, for these original datasets, the power losses are small for most OEPs, and some OEPs actually yield small improvements in power relative to no exclusions (i.e., fixed cutoffs, regular *Z* cutoffs, and trimming tend to reject HØ more than half of the time).

When they are applied to the original datasets in a pooled manner, the OEPs generally have more power than using no exclusions at all (i.e., they reject HØ more than half of the time—in some cases more than 60% of the time). These results provide some encouragement that an OEP applied in a pooled manner could improve statistical power.

For the transposed datasets with a stretch rather than shift word/nonword effect, however, almost all of the adaptive OEPs have clearly lower power than would be obtained with no exclusions, regardless of whether the OEP is applied with the condition-specific or pooled approach. The power loss is quite serious for some OEPs—especially when they are applied in a pooled manner—presumably because the adaptive OEPs are biased to underestimate the true effect sizes, as was predicted on theoretical grounds and as will be documented in the next section. It is thus easy to imagine that researchers have failed to detect many stretch-type effects as significant specifically because of the power loss associated with the OEP that they used. Furthermore, a clear pattern emerges when looking at transposed dataset power within each of the different types of OEPs (i.e., *Z*, simple recursive, MAD, trim, etc): For every type, power is worse when a lower cutoff criterion leads to exclusion of more RTs (e.g., a *Z* criterion of 2 rather than 2.5). That is, the more trials are excluded, the more likely researchers are to miss real effects that would be revealed by an analysis of the full dataset. This pattern is not surprising, because the evidence for a stretch effect would naturally be weakened by discarding more of the large RTs that show the effect most clearly, and that happens to a greater extent with the lower cutoffs (Figures 1 and 6).

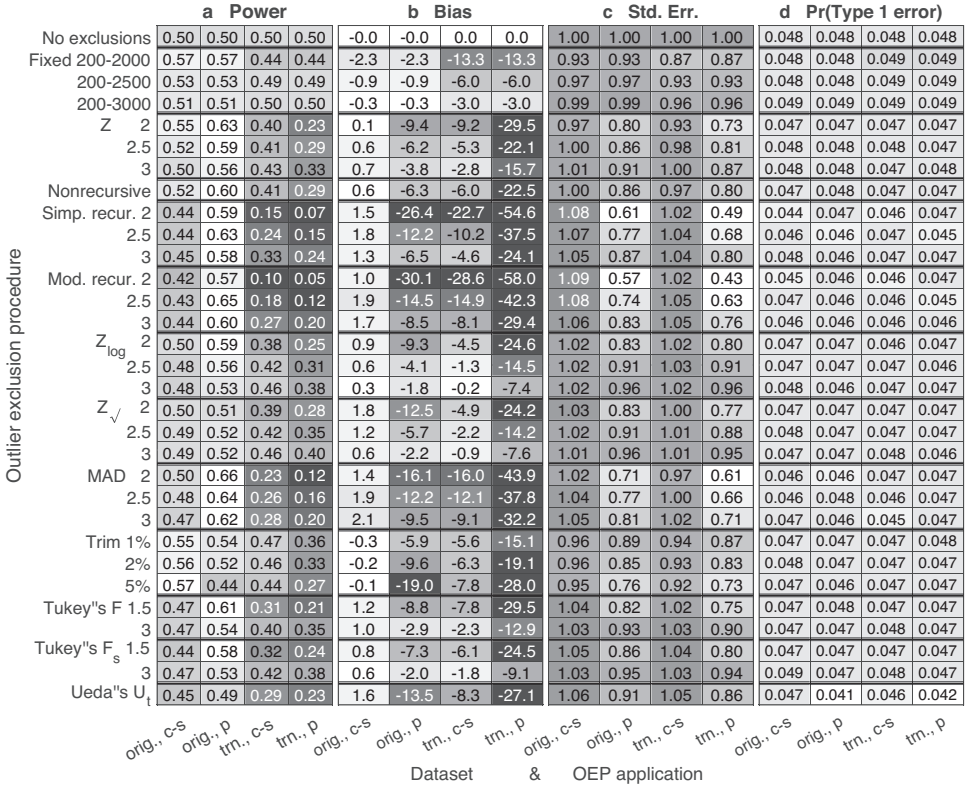
The original and transposed datasets examined here are extreme in the sense that they show entirely shift effects or entirely stretch effects, whereas many experimental manipulations produce intermediate datasets having each of these two types of effects to some degree (e.g., Rieger & Miller, 2020). With intermediate datasets, the power of the different OEPs would presumably be intermediate between the original and transposed power levels shown in Figure 15a. Given that all OEPs have power losses with the stretch effects that are larger than their power gains—if any—with shift effects, the overall conclusion from these simulations is that none of the OEPs should be used to increase power unless there is strong prior evidence of a pure shift effect.

Bias of Effect Size Measurement

In addition to assessing the statistical significance of an effect, researchers want to have unbiased estimates of the *size* of that effect (e.g., Fritz et al., 2012; Ulrich et al., 2018; Wilson et al., 2020). It is known, however, that effect size estimates can be biased by outlier exclusion, particularly with skewed distributions (e.g., Miller, 1991; Ulrich & Miller, 1994). Thus, it is also necessary to consider the possible biasing effects of each OEP.

Figure 15

Results of Simulations Investigating the Statistical Power (a), Bias in Effect Size Measurement (b), Standard Error of Effect Size Measurement (c), and Type 1 Error Rate (d) for Each Outlier Exclusion Procedure With Condition-Specific (c-s) or Pooled (p) Application, Separately for the Original (Orig.) and Transposed (Trn.) Datasets



Note. See the text for explanations of the measures shown in the four panels. For all measures, better performance (e.g., higher power, smaller bias) is indicated by lighter shading.

With the present technique of applying OEPs to real datasets rather than to data generated from prespecified distributions, the true effect size is unfortunately not known. Nonetheless, by using an assumed true effect size that is the same for all OEPs, it is at least possible to compare the *relative* bias introduced by the various OEPs. For the present analysis, the true effect size for each dataset was assumed to be the effect size observed in the analysis of all participants and RTs (i.e., no exclusions) for that dataset. Although these assumed effect sizes are not exactly correct, there are several reasons to suspect that they would be very close to the true values. First, deviations due to purely random error should be small due to the large numbers of participants and trials in each dataset. Second, the evidence already presented suggests that there are only small proportions of outliers in these datasets. Third, even if outliers caused by task-unrelated processing are present, they should affect the mean RTs for words and nonwords approximately equally, thereby having little or no effect on the RT difference between the two conditions (i.e., effect size).[11]

The bias values shown in Figure 15b depict the differences, in milliseconds, between the word/nonword effect sizes computed from the full datasets and the averages of the effect sizes computed across the 10,000 simulated experiments, again averaging across simulations with different numbers of participants and trials. These show that condition-specific application of the OEPs to the original datasets generally results in only small biases. With pooled application to the original datasets and with either type of application to the transposed datasets, however, effect sizes are nearly always underestimated, and the underestimations are substantial for many OEPs. For the transposed datasets, pooled application conceals approximately one-third to one-half of the true effect size in many cases (see Table 2). As was the case for power loss, the bias for each type of OEP is worse when the cutoff criterion leads to exclusion of more RTs, and the reason for this pattern is analogous: The estimated effect size tends to be reduced by excluding more of the slow RTs in which the stretching effect is most strongly present.

[11] One might imagine that task-unrelated processing happened more often in one condition than another; for example, there might be more distraction in the slower nonword condition. This should probably be considered part of the word/nonword condition difference, however, and not a contaminating effect of task-unrelated outliers. It certainly does not appear that this was an issue in the present datasets, because it would tend to produce a stretch effect at the high end of the nonword RT distributions, which was not observed.

Standard Error of Effect Size Measurement

The standard errors of the OEPs' effect size estimates are also of interest, because it is preferable to obtain estimates that are subject to lower levels of random variability (i.e., have smaller standard errors). Figure 15c shows the normalized standard errors of the effect size estimates obtained with each OEP. Naturally, before normalization, the raw standard errors were smaller in simulations with larger numbers of participants and trials per participant. To facilitate comparisons of the OEPs across conditions, the standard errors were normalized to reduce the influence of these numbers. To compute normalized standard errors, the raw standard error of the effect size for each OEP and simulation condition was computed as the standard deviation of the effect sizes across the simulated experiments. These raw values were then normalized for each OEP by taking the ratio of that OEP's raw standard error divided by the raw standard error obtained with the no-exclusion procedure in the same simulation condition (i.e., same dataset, number of participants, and number of trials). Thus, a normalized standard error greater than one indicates that an OEP's effect size estimates were more variable than the effect size estimates with no exclusions, whereas a normalized standard error less than one indicates that the OEP yielded effect size estimates that were less variable.

Relative to the no-exclusion approach, condition-specific OEP application has relatively small effects on the standard error of the effect size estimate, leading to normalized standard error values close to one. With condition-specific application, the fixed-cutoff, *Z*-score-based, and trimming OEPs produced slightly lower standard errors than using no exclusions, whereas the other OEPs produced slightly higher standard errors. On the other hand, pooled application substantially reduced the standard errors of the effect size estimates for many OEPs. With the original datasets, the reduction in standard error obtained with pooled application of many OEPs apparently outweighed the effect size reduction (i.e., bias) in order to maintain the good power seen in Figure 15a. With the transposed datasets, however, the standard error reductions were too small to compensate fully for the effect size reduction, leading to the serious power losses.

Type 1 Error Rate

Finally, a separate set of 2,500 simulations examining Type 1 error rate was run for each combination of dataset, OEP, number of participants, and number of trials. In these simulations, the HØ of no word/nonword difference was true by construction, because each participant's RTs for both conditions were randomly selected from the same distribution (i.e., the combined pool of all the word and nonword RTs from that participant). Thus, the proportion of significant results ($p < .05$) in these latter simulations is an estimate of the Type 1 error rate. As can be seen in Figure 15d, all of the OEPs performed very well on this measure, with Type 1 error rates just below the nominal 5% level that would be appropriate. This pattern was also found with simulated RTs for some of these OEPs by Ratcliff (1993). Finding Type 1 error rates at the expected level for the condition-specific approach is initially surprising given André's (2022) report that OEPs can inflate Type 1 error rates when applied separately to each condition. The hypothesis testing situation examined by André (2022) was, however, different from the present one in that it involved between-subjects designs with the possible exclusion of entire outlier participants rather than of individual trials within a condition. The contrast between André's results and those of the present simulations makes it clear that condition-specific application of OEPs has very different consequences for these two different types of designs.

Testing for Experimental Effects: Summary

Considering all of the simulation results relevant to testing for experimental effects in within-subjects designs, it is difficult to argue that researchers investigating such effects should regularly use *any* of the OEPs considered here. Some of the OEPs do produce modest increases in statistical power for detecting pure shift effects of the sort present in the original datasets, especially when they are applied in a pooled manner, although in these cases the effect sizes tend to be markedly underestimated. Unfortunately, these OEPs produce much larger decreases in power with stretch effects like that present in the transposed datasets, and in that case, they also introduce even greater biases to underestimate the true effect size. In no context did the OEPs consistently improve Type 1 error rates. Thus, especially when the investigated experimental comparison might involve a stretch effect, the potentially serious costs of using an OEP appear to outweigh any benefits.

If there were strong reasons to suspect the presence of many outliers in a dataset and therefore strong motivation to apply some OEP, the present results suggest that the best choice would be the $Z_{\sqrt{}}$ 3 or $Z_{\log}$ 3 OEP applied in a condition-specific manner. These OEPs suffer only small power losses and introduce only small biases in effect size estimation even with stretch effects. With the present datasets, these OEPs also tended to exclude among the fewest trials—just about 1%—but of course, they would presumably exclude higher percentages of trials from datasets with higher percentages of true outliers.

General Discussion

Each of the various OEPs considered in this article has a plausible theoretical rationale—perhaps especially the adaptive ones designed to identify outliers in a manner that allows for between-participant and possibly also between-condition differences. In practice, however, the choice of an OEP must be determined by examining the associated costs and benefits: Does an OEP identify outliers accurately, and does it thereby improve the statistical accuracy of condition comparisons?

The results of the present analyses suggest that both of these questions must be answered in the negative. I conclude that researchers testing for condition effects in typical within-subjects designs with reasonable maximum allowable RTs should not try to exclude RTs using any of these adaptive OEPs but should instead exclude only RTs that are obvious outliers within the context of their experimental tasks (i.e., using wide fixed cutoffs such as 150–3,000 ms, as examined by Ulrich & Miller, 1994). Conveniently, this conclusion also accords with one of the most fundamental tenets underlying statistical analysis: "All statisticians know that data are falsified if only a selected part is used…. Our conclusions must be warranted by the whole of the data, since less than the whole may be to any degree misleading" (Fisher, 1935, p. 54). An obvious corollary is that researchers should be generous in setting the maximum RT interval that is allowed for participants actually doing the task, since the length of this interval inherently limits the maximum observable RT. If researchers feel that an OEP must be applied despite the

problems documented here, they should justify that position by giving evidence that outliers are present in the first place and that the applied OEP identifies them correctly.

A major problem with both the condition-specific and pooled adaptive OEPs is that they are inconsistent in their outlier classifications. Classifications of the same RTs from the same participant and condition are inconsistent not only across analyses of RT samples with different numbers of trials but even across analyses of different random samples with the same number of trials. Even for a single participant in a single condition with a fixed and a reasonably large number of trials (e.g., 80–120), it appears that sample-to-sample variability is too great for accurate determination of the boundaries between inliers and outliers. As is illustrated in Figure 10a, for example, the *Z* 2.5 OEP might exclude an RT at the 94th percentile of the participant's RT distribution in one sample of 80 trials and yet not exclude an RT at the 98th percentile of that distribution in a different sample of 80 trials. Such inconsistencies are present for all of the adaptive OEPs, regardless of whether their boundaries are computed from parametric (e.g., *Z*-scores of the RTs or of their transformed values) or nonparametric (e.g., MAD scores, percentiles, Tukey's fences) measures. The inconsistencies are greater with smaller numbers of trials, and they show that the classifications cannot possibly all be accurate with respect to each RT's true inlier/outlier status in real datasets. The inconsistencies are slightly smaller with pooled than condition-specific application of the OEPs, presumably because the cutoff boundaries are more stable when they are computed from the larger numbers of trials examined with the pooled approach, but they remain large enough in absolute terms to raise serious doubts about the use of these OEPs.

Admittedly, it might still be worthwhile to use some OEP—despite its less than perfect accuracy—if outliers had very harmful consequences for the main analyses. This seems unlikely in principle, however. Since outliers arise from task-unrelated processes, they should occur with the same probabilities in all conditions and have the same influences on all condition mean RTs. In that case they would not bias the between-condition differences in means but would only add noise to the comparisons.

The present analyses suggest that outliers were rare in the present datasets and that the noise added by attempts to exclude them is greater than the noise tolerated by retaining them. It seems clear that the outliers were rare, because the percentages of RTs identified as outliers in the real datasets were nearly identical to the percentages of RTs identified as outliers in outlier-free datasets (i.e., the ØO datasets generated from ex-Gaussian distributions), and this was true for each of the OEPs. Yet, some of the OEPs have nontrivial base rates of classifying valid RTs as outliers—that is, nontrivial FP rates. Further evidence of this can be seen in the analysis of subsamples of RTs, where outlier classification FPs were common even relative to the classifications of the same RTs by the same OEPs in the full samples of RTs (Figure 7). Given this evidence of OEP FPs, it is doubtful that many of the to-be-excluded RTs were actually outliers (i.e., caused by task-unrelated processes) despite the fact that some of the OEPs classified them as such.

With respect to the effectiveness of the OEPs in removing noise from experimental comparisons, the results summarized in Figure 15 paint a disturbing picture. Although some of the OEPs perform better than the no-exclusions procedure under certain conditions, none of the OEPs is consistently as good as or better than the procedure of simply analyzing all of the RTs with no exclusions. For example, many of the OEPs introduce a substantial underestimation bias, especially when testing for stretch effects that manifest primarily in the slowest responses. Furthermore, although it might intuitively seem that OEPs would help prevent outliers from causing Type 1 errors, the simulations provided no evidence of such an effect (Figure 15d).

With respect to the key issue of power, the present results show that some OEPs can increase it for the detection of shift effects, relative to the no-exclusions procedure. Berger and Kiefer (2021) reached the same conclusion based on their simulations with artificially generated RTs, all of which involved only pure shift effects. Unfortunately, the results also show that the OEPs can cause much greater decreases in power with stretch effects. The power loss with stretch effects can be large even when the percentage of slow RTs eliminated is rather small (e.g., simple recursive 3), presumably because the eliminated slow RTs are the ones most affected by the stretch. This power loss is especially large with pooled OEP application because this approach is especially likely to exclude the most-affected slow RTs. Thus, researchers would be well advised to forego these OEPs to avoid the risk of power loss in experimental comparisons that have even the potential to produce a stretch effect. It is ironic that the OEPs perform especially badly with the greater skew found in the case of stretch effects, because RT skew is one of the main signs of the outlier problem that OEPs were intended to solve.

These results emphasize the importance of researchers providing information about the stretch versus shift nature of any new effect that is found. In the past, this has usually been done by separating the manipulation's effects on the μ and τ parameters of fitted ex-Gaussian distributions, but it could be done more simply by reporting effect sizes separately for RTs faster versus slower than each participant's condition median RT. Shift effects would increase the faster-than-median and slower-than-median RTs to about the same degree, whereas stretch effects would increase the slower ones much more than the faster ones.

Implications for Prior Research?

The current results suggest that many valid RTs in previous studies would incorrectly have been excluded as outliers by the OEPs that were in use. What, then, are the implications of these likely incorrect previous exclusions for the accuracy of previous findings (e.g., Type 1 error rates, effect size estimates, replicability)?

Unfortunately, answering this question in detail requires very specific assumptions about the underlying research context (e.g., base rate of true effects, distribution of effect sizes, etc.; Miller & Schwarz, 2011; Miller & Ulrich, 2016), but some general points can be made. First, as shown in Figure 15b, previous studies may have tended to underestimate true effect sizes because of the OEPs that were used—especially if the comparisons involved stretching of the RT distribution in the slower condition. On the plus side, the tendency of OEPs to reduce estimated effect size could actually have increased the final accuracy of published effect size estimates by counteracting the well-known overestimation of effect size bias associated with the selection of statistically significant effects (e.g., Ulrich et al., 2018). However, it is scientifically slipshod to rely on compensating biases to arrive at correct values.

Second, given that the OEPs do not inflate Type 1 error rates (Figure 15d), there might initially seem to be no increased risk of Type 1 errors among published studies using OEPs. Unfortunately, this viewpoint is overly optimistic. For one thing, if the study was not preregistered then there is always the possibility

that the reported OEP is just one of several that were tried in attempts to get "clean" data (i.e., $p < .05$), and it is well known that trying multiple OEPs tends to inflate the Type 1 error rate in RT analyses (Berger & Kiefer, 2021; Morís Fernández & Vadillo, 2020; Ulrich & Miller, 2020). For another, the probability that a significant effect ($p < .05$) is a Type 1 error tends to be larger among studies that have less power, for purely statistical reasons (e.g., Ioannidis, 2005). Thus, the tendency of the OEPs to reduce power under at least some circumstances (Figure 15a) indirectly increases the proportion of published results that are expected to be Type 1 errors.

Caveats

Although the present analyses using real RTs provide a useful supplement to previous simulations based on specific distributional assumptions, they do come with several limitations of their own. First, this analysis pertains only to within-subject comparisons of the mean RTs in two different experimental conditions. It is still an open question whether these same OEPs would be helpful with other experimental designs (e.g., between-subjects comparisons). It is also possible, at least in principle, that the OEPs could be helpful with measures other than mean RTs (e.g., delta plots and estimates of RT model parameters). This seems unlikely, however, given the clear signs that the OEPs exclude extreme yet valid RTs, which could have serious consequences for RT measures that are sensitive to full RT distributions.

Second, the analysis was limited to studies of normal adult participants performing lexical decision tasks with reasonable limits on the maximum allowable RT. It is easy to imagine that the percentage of outliers would be higher with other tasks or participant groups (e.g., children; Leth-Steensen et al., 2000) or that outliers might be more troublesome without such limits on the allowable RT, and in such situations OEPs might be very helpful. The present results suggest, however, that the routine application of OEPs without clear evidence of outliers is not an optimal analytic approach in typical RT tasks. The lexical decision task may also be somewhat extreme with respect to the heterogeneity of its stimulus materials (cf., Figure 4's item analyses), but many tasks use varying stimulus displays (e.g., Beanland et al., 2016; Sanocki et al., 2015) and most involve trial-to-trial difficulty fluctuations caused by factors such as stimulus and response recency and repetition effects (e.g., Bertelson, 1965; Rabbitt & Vyas, 1979). Similarly, the relative performance of the different OEPs might change with other differences between conditions at the level of the RT distribution. Both shifting and stretching effects were examined here, but there are actually other possibilities; for example, with some manipulations, the slower condition tends to be *less* skewed (i.e., have a shorter upper tail) than the faster one (e.g., Schwarz & Miller, 2012).

Third, the present analysis was limited to the OEPs shown in Table 1. Although this is a relatively complete representation of the OEPs currently available, it is certainly possible that researchers will devise new or improved OEPs that could be used instead of the ones examined here, because there is an ongoing development of alternative methods of outlier identification within applied statistics (e.g., Cafaro et al., 2021; Hubert & Van der Veeken, 2008; Rousseeuw & Hubert, 2011). The present analyses suggest some ways in which any proposed new OEPs could be evaluated using real data to assess the accuracy and stability of their outlier classifications and their effects on power and bias in detecting and measuring condition effect sizes. More generally, these results emphasize that the selection of optimal data analysis procedures is an empirical issue and that it is risky to choose procedures based solely on a compelling intuitive rationale.

References

Alexandrowicz, R. W. (2020). The diffusion model visualizer: An interactive tool to understand the diffusion model parameters. *Psychological Research*, *84*(4), 1157–1165. https://doi.org/10.1007/s00426-018-1112-6

André, Q. (2022). Outlier exclusion procedures must be blind to the researcher's hypothesis. *Journal of Experimental Psychology: General*, *151*(1), 213–223. https://doi.org/10.1037/xge0001069

Aucremanne, L., Brys, G., Hubert, M., Rousseeuw, P. J., & Struyf, A. (2004). A study of Belgian inflation, relative prices and nominal rigidities using new robust measures of skewness and tail weight. In M. Hubert, G. Pison, A. Struyf, & S. Van Aelst (Eds.), *Theory and applications of recent robust methods* (pp. 13–25). Birkhauser. https://doi.org/10.1007/978-3-0348-7958-3_2

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples *t* tests: The power of alternatives and recommendations. *Psychological Methods*, *19*(3), 409–427. https://doi.org/10.1037/met0000014

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Baykara, S., & Alban, K. (2019). Visual and auditory reaction times of patients with opioid use disorder. *Psychiatry Investigation*, *16*(8), 602–606. https://doi.org/10.30773/pi.2019.05.16

Beanland, V., Le, R. K., & Byrne, J. E. M. (2016). Object–scene relationships vary the magnitude of target prevalence effects in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(6), 766–775. https://doi.org/10.1037/xhp0000183

Beesley, T., Vadillo, M. A., Pearson, D., & Shanks, D. R. (2016). Configural learning in contextual cuing of visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1173–1185. https://doi.org/10.1037/xhp0000185

Berger, A., & Kiefer, M. (2021). Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology*, *12*, Article 675558. https://doi.org/10.3389/fpsyg.2021.675558

Bertelson, P. (1965). Serial choice reaction-time as a function of response versus signal-and-response repetition. *Nature*, *206*(4980), 217–218. https://doi.org/10.1038/206217a0

Brosowsky, N. P., & Egner, T. (2021). Appealing to the cognitive miser: Using demand avoidance to modulate cognitive flexibility in cued and voluntary task switching. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(10), 1329–1347. https://doi.org/10.1037/xhp0000942

Burke, D., & Roodenrys, S. (2000). Implicit learning in a simple cued reaction-time task. *Learning and Motivation*, *31*(4), 364–380. https://doi.org/10.1006/lmot.2000.1062

Bush, L. K., Hess, U., & Wolford, G. (1993). Transformations for within-subject designs: A Monte Carlo investigation. *Psychological Bulletin*, *113*(3), 566–579. https://doi.org/10.1037/0033-2909.113.3.566

Cafaro, M., Melle, C., Pulimeno, M., & Epicoco, I. (2021). Fast online computation of the Qn estimator with applications to the detection of outliers in data streams. *Expert Systems with Applications*, *164*, Article 113831. https://doi.org/10.1016/j.eswa.2020.113831

Clithero, J. A. (2018). Response times in economics: Looking through the lens of sequential sampling models. *Journal of Economic Psychology*, *69*, 61–86. https://doi.org/10.1016/j.joep.2018.09.008

Cochrane, B. A., & Pratt, J. (2022). The item-specific proportion congruency effect can be contaminated by short-term repetition priming. *Attention,*

Perception, and Psychophysics, *84*(1), 1–9. https://doi.org/10.3758/s13414-021-02403-0

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research*, *3*(1), 58–67. https://doi.org/10.21500/20112084.844

Donders, F. C. (1868/1969). Over de snelheid van psychische processen. [On the speed of mental processes.] (W. G. Koster, Trans.). In W. G. Koster (Ed.), *Attention and performance II* (pp. 412–431). North Holland. https://doi.org/10.1016/0001-6918(69)90065-1

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception and Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496. https://doi.org/10.3758/BRM.42.2.488

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*(1), 39–82. https://doi.org/10.2307/2342435

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18. https://doi.org/10.1037/a0024338

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response-time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*(2), 340–347. https://doi.org/10.1037/0033-2909.109.2.340

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 598–615. https://doi.org/10.1037/0278-7393.10.4.598

Hockley, W. E., & Corballis, M. C. (1982). Tests of serial scanning in item recognition. *Canadian Journal of Psychology*, *36*(2), 189–212. https://doi.org/10.1037/h0080637

Hoedemaker, R. S., & Gordon, P. C. (2017). The onset and time course of semantic priming during rapid recognition of visual words. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(5), 881–902. https://doi.org/10.1037/xhp0000377

Hubert, M., & Van der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics*, *22*(3–4), 235–246. https://doi.org/10.1002/cem.1123

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, *52*(12), 5186–5201. https://doi.org/10.1016/j.csda.2007.11.008

Hübner, R., Töbel, L., & Kumari, V. (2019). Conflict resolution in the Eriksen flanker task: Similarities and differences to the Simon task. *PLoS ONE*, *14*(3), e0214203. https://doi.org/10.1371/journal.pone.0214203

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., Yap, M. J., Bengson, J. J., Niemeyer, D., & Buchanan, E. (2013). The Semantic Priming Project. *Behavior Research Methods*, *45*(4), 1099–1114. https://doi.org/10.3758/s13428-012-0304-z

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Janczyk, M., & Ulrich, R. (2019). Action consequences affect the space-time congruency effect on reaction time. *Acta Psychologica*, *198*, Article 102850. https://doi.org/10.1016/j.actpsy.2019.05.002

Johnston, C. D., & Madson, G. J. (2022). Negativity bias, personality and political ideology. *Nature Human Behaviour*, *6*(5), 666–676. https://doi.org/10.1038/s41562-022-01327-5

Karch, J. D. (2023). Outliers may not be automatically removed. *Journal of Experimental Psychology: General*, *152*(6), 1735–1753. https://doi.org/10.1037/xge0001357

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578–589. https://doi.org/10.1037/met0000209

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, *1*, Article 174. https://doi.org/10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

Kimber, A. C. (1990). Exploratory data analysis for possibly censored data from skewed distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *39*(1), 21–30. https://doi.org/10.2307/2347808

Krummenacher, J., Müller, H. J., & Heller, D. (2002). Visual search for dimensionally redundant pop-out targets: Redundancy gains in compound tasks. *Visual Cognition*, *9*(7), 801–837. https://doi.org/10.1080/13506280143000269

Leth-Steensen, C., Elbaz, Z. K., & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: A response time distributional approach. *Acta Psychologica*, *104*(2), 167–190. https://doi.org/10.1016/S0001-6918(00)00019-6

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, *32*(1), Article 5. https://doi.org/10.5334/irsp.289

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Lohani, M., Cooper, J. M., Erickson, G. G., Simmons, T. G., McDonnell, A. S., Carriero, A. E., Crabtree, K. W., & Strayer, D. L. (2021). No difference in arousal or cognitive demands between manual and partially automated driving: A multi-method on-road study. *Frontiers in Neuroscience*, *15*, Article 627. https://doi.org/10.3389/fnins.2021.577418

Low, K. A., Miller, J. O., & Vierck, E. (2002). Response slowing in Parkinson's disease: A psychophysiological analysis of premotor and motor processes. *Brain*, *125*(9), 1980–1994. https://doi.org/10.1093/brain/awf206

Lubczyk, T., Lukács, G., & Ansorge, U. (2022). Speed versus accuracy instructions in the response time concealed information test. *Cognitive Research: Principles and Implications*, *7*(1), Article 3. https://doi.org/10.1186/s41235-021-00352-8

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

Marmolejo-Ramos, F., Vélez, J. I., & Romão, X. (2015). Automatic detection of discordant outliers via the Ueda's method. *Journal of Statistical Distributions and Applications*, *2*(1), Article 8. https://doi.org/10.1186/s40488-015-0031-y

Mazor, M., & Fleming, S. M. (2022). Efficient search termination without task experience. *Journal of Experimental Psychology: General*, *151*(10), 2494–2510. https://doi.org/10.1037/xge0001188

Miller, J. O. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The Quarterly Journal of Experimental Psychology Section A*, *43*(4), 907–912. https://doi.org/10.1080/14640749108400962

Miller, J. O. (2021). Percentile rank pooling: A simple nonparametric method for comparing group reaction time distributions with few trials. *Behavior Research Methods*, *53*(2), 781–791. https://doi.org/10.3758/s13428-020-01466-5

Miller, J. O., & Schwarz, W. (2011). Aggregate and individual replication probability within an explicit model of the research process. *Psychological Methods*, *16*(3), 337–360. https://doi.org/10.1037/a0023347

Miller, J. O., & Tang, J. L. (2021). Effects of task probability on prioritized processing: Modulating the efficiency of parallel response selection.

Attention, Perception, and Psychophysics, *83*(1), 356–388. https://doi.org/10.3758/s13414-020-02143-7

Miller, J. O., & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science*, *11*(5), 664–691. https://doi.org/10.1177/1745691616649170

Morís Fernández, L., & Vadillo, M. A. (2020). Flexibility in reaction time analysis: Many roads to a false positive? *Royal Society Open Science*, *7*(2), Article 190831. https://doi.org/10.1098/rsos.190831

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), 367–374. https://doi.org/10.1037/a0025858

Mowrer, O. H., Rayman, N., & Bliss, E. (1940). Preparatory set (expectancy)—An experimental demonstration of its "central" locus. *Journal of Experimental Psychology*, *26*(4), 357–372. https://doi.org/10.1037/h0058172

Myers, C. E., Interian, A., & Moustafa, A. A. (2022). A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences. *Frontiers in Psychology*, *13*, Article 1039172. https://doi.org/10.3389/fpsyg.2022.1039172

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), 58–71. https://doi.org/10.1037/a0020747

Paylor, R. (2009). Questioning standardization in science. *Nature Methods*, *6*(4), 253–254. https://doi.org/10.1038/nmeth0409-253

Pesce, C., & Audiffren, M. (2011). Does acute exercise switch off switch costs? A study with younger and older athletes. *Journal of Sport and Exercise Psychology*, *33*(5), 609–626. https://doi.org/10.1123/jsep.33.5.609

Plater, L., Giammarco, M., Fiacconi, C., & Al-Aidroos, N. (2020). No role for activated long-term memory in attentional control settings. *Journal of Experimental Psychology: General*, *149*(2), 209–221. https://doi.org/10.1037/xge0000642

Possamaï, C. A. (1991). A responding hand effect in a simple-RT precueing experiment: Evidence for a late locus of facilitation. *Acta Psychologica*, *77*(1), 47–63. https://doi.org/10.1016/0001-6918(91)90064-7

Rabbitt, P. M. A., & Vyas, S. M. (1979). Signal recency effects can be distinguished from signal repetition effects in serial CRT tasks. *Canadian Journal of Psychology*, *33*(2), 88–95. https://doi.org/10.1037/h0081708

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Rieger, T. C., & Miller, J. O. (2020). Are model parameters linked to processing stages? An empirical investigation for the ex-Gaussian, ex-Wald, and EZ diffusion models. *Psychological Research*, 84(6), 1683–1699. https://doi.org/10.1007/s00426-019-01176-4

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *WIRES Data Mining and Knowledge Discovery*, *1*(1), 73–79. https://doi.org/10.1002/widm.2

Sali, A. W., Ma, R., Albal, M. S., & Key, J. (2022). The location independence of learned attentional flexibility. *Attention, Perception, and Psychophysics*, *84*(3), 682–699. https://doi.org/10.3758/s13414-022-02469-4

Sanocki, T., Islam, M., Doyon, J. K., & Lee, C. (2015). Rapid scene perception with tragic consequences: Observers miss perceiving vulnerable road users, especially in crowded traffic scenes. *Attention, Perception, and Psychophysics*, *77*(4), 1252–1262. https://doi.org/10.3758/s13414-015-0850-4

Schwarz, W., & Miller, J. O. (2012). Response time models of delta plots with negative-going slopes. *Psychonomic Bulletin and Review*, *19*(4), 555–574. https://doi.org/10.3758/s13423-012-0254-6

Seow, T., & Fleming, S. M. (2019). Perceptual sensitivity is modulated by what others can see. *Attention, Perception, and Psychophysics*, *81*(6), 1979–1990. https://doi.org/10.3758/s13414-019-01724-5

Shimizu, Y. (2022). Multiple desirable methods in outlier detection of univariate data with R source codes. *Frontiers in Psychology*, *12*, Article 819854. https://doi.org/10.3389/fpsyg.2021.819854

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Singh, T., Laub, R., Burgard, J. P., & Frings, C. (2018). Disentangling inhibition-based and retrieval-based aftereffects of distractors: Cognitive versus motor processes. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(5), 797–805. https://doi.org/10.1037/xhp0000496

Steinhauser, M., & Hübner, R. (2009). Distinguishing response conflict and task conflict in the Stroop task: Evidence from ex-Gaussian distribution analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 35(5), 1398–1412. https://doi.org/10.1037/a0016467

Tukey, J. W. (1959). A quick, compact, two-sample test to Duckworth's specifications. *Technometrics*, *1*(1), 31–48. https://doi.org/10.1080/00401706.1959.10489847

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Ueda, T. (2009). A simple method for the detection of outliers. *Electronic Journal of Applied Statistical Analysis*, *2*(1), 67–76. [(F. Marmolejo-Ramos & S. Kinoshita, Trans.) (Original work published in 1996)]. https://doi.org/10.1285/i20705948v2n1p67

Ulrich, R., & Miller, J. O. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, *123*(1), 34–80. https://doi.org/10.1037/0096-3445.123.1.34

Ulrich, R., & Miller, J. O. (2020). Meta-research: Questionable research practices may have little effect on replicability. *eLife*, *9*, Article e58237. https://doi.org/10.7554/eLife.58237

Ulrich, R., Miller, J. O., & Erdfelder, E. (2018). Effect size estimation from *t*-statistics in the presence of publication bias: A brief review of existing approaches with some extensions. *Zeitschrift für Psychologie*, *226*(1), 56–80. https://doi.org/10.1027/2151-2604/a000319

Van Selst, M., & Jolicœur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section A*, *47*(3), 631–650. https://doi.org/10.1080/14640749408401131

Wang, H., Walenski, M., Litcofsky, K., Mack, J. E., Mesulam, M. M., & Thompson, C. K. (2022). Verb production and comprehension in primary progressive aphasia. *Journal of Neurolinguistics*, *64*, Article 101099. https://doi.org/10.1016/j.jneuroling.2022.101099

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences*, *117*(11), 5559–5567. https://doi.org/10.1073/pnas.1914237117

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79. https://doi.org/10.1037/a0024177

Zhou, B., & Krott, A. (2016). Data trimming procedure can eliminate bilingual cognitive advantage. *Psychonomic Bulletin and Review*, *23*(4), 1221–1230. https://doi.org/10.3758/s13423-015-0981-6

(*Appendices follow*)

Appendix A

Construction of Transposed Datasets

This appendix describes the computational procedure used to construct the four transposed datasets with the word/nonword effect changed from the shift effect that was present in the original data to a stretch effect of the sort sometimes seen with other experimental manipulations. The presence of a shift effect in the original RTs will be documented first, and then the procedure for generating nonword RTs exhibiting a stretch effect will be outlined.

One common way to describe the effects of experimental manipulations on RT distributions is to summarize each observed RT distribution (i.e., for a given participant in a given condition) with the μ, σ, and τ parameters of the best-fitting ex-Gaussian (e.g., Heathcote et al., 1991). Between-condition differences in μ basically reflect a shift in the distribution, whereas differences in τ reflect differences in the stretch or skewness in the upper tail of the RT distribution.

Table A1 summarizes the original data's word/nonword effect in this way, showing the means across participants of the ex-Gaussian μ, σ, and τ parameters estimated separately for each participant and condition by maximum likelihood. These are fairly typical of ex-Gaussian parameter values also seen in other tasks, with the Gaussian component μ accounting for approximately 70% of the mean RT, and the Gaussian variance σ^2 accounting for only about 5%–10% of the total RT variance $\sigma^2 + \tau^2$. Critically, it is evident that in these datasets almost all of the mean RT difference between the word and nonword conditions is in μ rather than τ—that is, the word/nonword difference represents a shift effect.

To broaden the conclusions from the present analyses, it was desirable to produce a dataset that was as close to the real dataset as possible but with the word/nonword effect "transposed" from the μ to the τ parameter of the ex-Gaussian, thus producing realistic datasets with a stretch effect rather than a shift effect. The following ad hoc procedure was developed for that purpose, and it was carried out separately for each participant in each dataset. Only the participants' RTs in the nonword condition were changed for the transposed datasets relative to the original ones; the word RTs were not altered.

For each participant, the first step was to estimate the ex-Gaussian parameters μ_W, μ_N, τ_W, and τ_N for the word and nonword conditions; these are the same estimated parameter values summarized in Table A1.

Next, for each RT in the nonword condition *N*, the value of the cumulative distribution function (CDF) $F_N(\mathrm{RT})$ was computed relative to the participant's estimated nonword ex-Gaussian (i.e., the ex-Gaussian with its estimated values of μ_N, σ_N and τ_N). Across the set of all of the participant's nonword RTs, this yields a set of individual-trial cumulative probability $F_N(\mathrm{RT})$ values, each ranging from zero to one. To the extent that the original nonword distribution was accurately described by the ex-Gaussian, these $F_N(\mathrm{RT})$ values should be uniformly distributed. Note, however, that if the inverse CDF transformation F^{-1} were applied to these $F_N(\mathrm{RT})$ values, then the original nonword RTs would be produced again exactly, regardless of whether the original nonword RT distribution was actually ex-Gaussian.

Next, a new transposed ex-Gaussian distribution (*N′*) was formed for each participant's nonword condition, transposing the participant's effect on μ to an effect on τ, and vice versa. More precisely, the new μ'_N and τ'_N for the transposed nonword ex-Gaussian distribution were formed from the participant's original estimates of μ and τ in each condition as

$$\mu'_N = \mu_W + (\tau_N - \tau_W)\,. \tag{A1}$$

$$\tau'_N = \tau_W + (\mu_N - \mu_W). \tag{A2}$$

Thus, the participant's original word/nonword effect on μ was transposed to an effect on τ, and vice versa. For each participant, this defines a transposed nonword ex-Gaussian distribution with the new μ'_N and τ'_N parameters; the original estimated σ_N was used again for the transposed distribution.

Finally, each of the participant's nonword RTs in the transposed condition, RT'_N, was computed as the inverse CDF $F_{N'}^{-1}$ of the original CDF value $F_N(\mathrm{RT})$, taking the inverse CDF relative to the transposed nonword distribution with parameters μ'_N and τ'_N:

$$\mathrm{RT}_{N'} = F_{N'}^{-1}[F_N(\mathrm{RT})]. \tag{A3}$$

Table A1
Mean Estimates of the Ex-Gaussian Parameters μ, σ, and τ in the Word (W) and Nonword (N) Conditions, and the Nonword Minus Word Differences (D) in μ and τ as Measures of Effect Size

	Dataset			
Parameter	SPP	DLP	ELP	FLP
μ_W	491.8	480.1	507.4	542.4
μ_N	555.1	508.7	586.5	629.5
σ_W	50.6	44.2	62.0	48.8
σ_N	53.3	48.3	76.7	67.6
τ_W	189.5	166.4	286.1	207.3
τ_N	197.6	167.6	294.8	216.8
μ_D	63.3	28.7	79.1	87.1
τ_D	8.1	1.2	8.6	9.5

Note. SPP = Semantic Priming Project; DLP = Dutch Lexicon Project; ELP = English Lexicon Project; FLP = French Lexicon Project.

(*Appendices continue*)

Note that any deviations from the ex-Gaussian distribution of the original nonword RTs, which would be reflected in their nonuniform set of F_N(RT) values, would be reproduced in the transposed RTs as deviations from the uniform with respect to the transposed ex-Gaussian. Thus, the procedure does not require the assumption that the true underlying distributions are exactly ex-Gaussian, but only the assumption that deviations from the ex-Gaussian will be correspondingly represented in the transposed distribution via the nonuniformity of the F_N(RT) values.

As a numerical example, consider a participant for whom the estimated ex-Gaussian parameters in the original dataset were $\mu_W = 400$, $\sigma_W = 40$, $\tau_W = 50$, $\mu_N = 450$, $\sigma_N = 45$, and $\tau_N = 55$ in the word and nonword conditions, respectively. The mean RTs in the two conditions are thus 450 and 505 ms, for an overall effect size of 55 ms. Using Equations A1 and A2, the parameters for this participant's transposed nonword condition would be $\mu'_N = 405$, $\sigma'_N = 45$, and $\tau'_N = 100$, so the mean of the nonword RTs would still be 505 ms. Suppose further that the first three original nonword RTs for this participant were 500, 550, and 600 ms. The CDFs of these three RTs within the original nonword ex-Gaussian are 0.52, 0.78, and 0.91, respectively. The RTs having those same three CDF values within the transposed nonword distribution are 487.1, 565.7, and 654.5, so these would be the first three RTs in the transposed nonword condition for this participant. As this example illustrates, for a participant whose word/nonword difference is mostly in μ, nonword RTs at the high end of the distribution are much slower after transposition than they were in the original nonword condition, whereas those at the low end of the distribution tend to be a bit faster after transposition.

Appendix B

Median *p* Values Used as α Levels

Table B1 shows the median p values obtained with the t-tests comparing words versus nonwords across the 10,000 simulations in which the RTs were analyzed with no exclusions. Medians are shown separately for each combination of the dataset, the number of participants, and the number of trials per word/nonword condition used in the simulation. These median p values were used as the α

Table B1
Median p *Values in Analyses With No Exclusions*

	N trials						
Dataset and *N* participants	10	20	40	60	80	120	160
Original dataset, 10 participants							
SPP	0.0582	0.0198	0.00654	0.00349	0.00235	0.00146	0.00109
DLP	0.272	0.172	0.103	0.0753	0.061	0.047	0.0385
ELP	0.0895	0.0331	0.0118	0.00634	0.00414	0.00251	0.0019
FLP	0.0276	0.01	0.00438	0.00289	0.00236	0.00184	0.00173
Original dataset, 20 participants							
SPP	0.00779	0.00108	0.000121	3.50E−05	1.74E−05	6.41E−06	3.64E−06
DLP	0.146	0.0574	0.0206	0.0117	0.0082	0.00521	0.00396
ELP	0.0188	0.00324	0.00041	0.000115	5.08E−05	2.00E−05	1.09E−05
FLP	0.00175	0.000267	4.45E−05	2.24E−05	1.42E−05	9.20E−06	7.27E−06
Original dataset, 40 participants							
SPP	0.000209	4.40E−06	6.14E−08	6.29E−09	1.27E−09	2.24E−10	6.95E−11
DLP	0.0413	0.00855	0.00136	0.000446	0.000216	9.21E−05	5.36E−05
ELP	0.00108	3.44E−05	6.11E−07	5.83E−08	1.26E−08	1.87E−09	5.09E−10
FLP	1.04E−05	2.08E−07	7.51E−09	2.04E−09	7.79E−10	3.20E−10	1.92E−10
Transposed dataset, 10 participants							
SPP	0.0857	0.0311	0.00971	0.00478	0.00324	0.00187	0.00132
DLP	0.307	0.203	0.116	0.0845	0.0646	0.0519	0.0427
ELP	0.124	0.0487	0.0169	0.00821	0.0056	0.00312	0.00216
FLP	0.0452	0.0158	0.00596	0.00378	0.00301	0.00212	0.00178
Transposed dataset, 20 participants							
SPP	0.0149	0.00222	0.000243	6.81E−05	3.12E−05	9.97E−06	5.27E−06
DLP	0.166	0.072	0.0274	0.0149	0.0096	0.00585	0.00432
ELP	0.0322	0.00591	0.000727	0.000224	9.36E−05	3.07E−05	1.59E−05
FLP	0.00485	0.000673	0.000104	4.23E−05	2.20E−05	1.28E−05	9.07E−06
Transposed dataset, 40 participants							
SPP	0.000715	1.96E−05	2.58E−07	2.23E−08	4.05E−09	5.28E−10	1.57E−10
DLP	0.0582	0.0131	0.002	0.000654	0.000336	0.000126	7.03E−05
ELP	0.00255	0.000114	2.09E−06	1.84E−07	3.64E−08	4.03E−09	1.07E−09
FLP	6.29E−05	1.29E−06	3.22E−08	6.49E−09	2.06E−09	6.12E−10	3.19E−10

Note. SPP = Semantic Priming Project; DLP = Dutch Lexicon Project; ELP = English Lexicon Project; FLP = French Lexicon Project.

(*Appendices continue*)

cutoffs for deciding whether to reject HØ in simulated experiments using all of the OEPs. For example, in a simulated experiment with the original SPP dataset, 20 participants, and 20 trials, the HØ was rejected for any OEP if its *t*-test yielded $p < .00108$.

Received September 12, 2022
Revision received April 4, 2023
Accepted May 18, 2023 ■