ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2020, Vol. 15(5) 1158–1177
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1745691620917678
www.psychologicalscience.org/PPS
SAGE

The Multiverse of Methods: Extending the Multiverse Analysis to Address Data-Collection Decisions

Jenna A. Harder
Department of Psychology, Michigan State University

Abstract

When analyzing data, researchers may have multiple reasonable options for the many decisions they must make about the data—for example, how to code a variable or which participants to exclude. Therefore, there exists a *multiverse* of possible data sets. A classic multiverse analysis involves performing a given analysis on every potential data set in this multiverse to examine how each data decision affects the results. However, a limitation of the multiverse analysis is that it addresses only data cleaning and analytic decisions, yet researcher decisions that affect results also happen at the data-collection stage. I propose an adaptation of the multiverse method in which the multiverse of data sets is composed of real data sets from studies varying in data-collection methods of interest. I walk through an example analysis applying the approach to 19 studies on shooting decisions to demonstrate the usefulness of this approach and conclude with a further discussion of the limitations and applications of this method.

Keywords

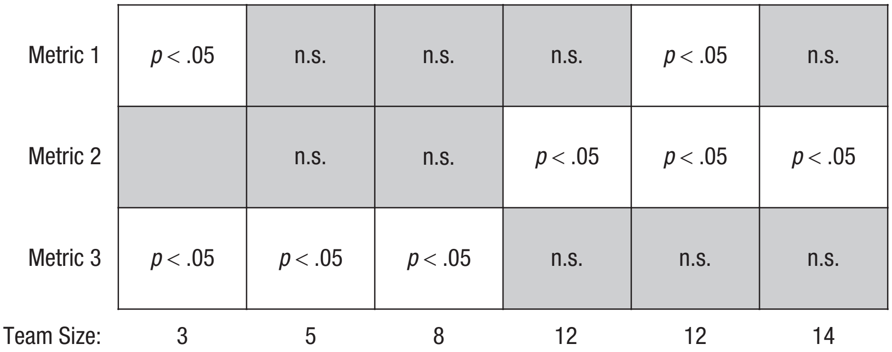
multiverse analysis, best practice, shooter bias, researcher degrees of freedom, multilevel modeling

The *multiverse analysis*, as termed by Steegen, Tuerlinckx, Gelman, and Vanpaemel (2016), is premised on the idea that in every analysis there are multiple reasonable options for the many decisions researchers must make about the data they have collected. These decisions could include rules for which participants are excluded, how certain variables are operationalized (e.g., what denotes the onset of puberty for male adolescents), how models are specified, and other judgment calls. Therefore, there exists for each analysis a set of potential data sets and analyses that could have been used instead—in other words, a “multiverse” of data sets and analyses. In a classic multiverse analysis, a single data set is collected or simulated, and a multiverse of data sets is generated by performing every possible combination of data-cleaning decisions. After this step is completed, analyses are performed on every potential data set in the multiverse to assess the extent to which each data decision affects the significance of the results.

The purpose of this method is to address researcher degrees of freedom by making the consequence of each of these choices transparent and to detect which choices have true implications for study conclusions. The multiverse analysis does not provide any information about the correct option in each of these choices. Rather, it provides descriptive information demonstrating the sensitivity of analyses to each data or analytic decision. The multiverse analysis can thus indicate the robustness of an existing study’s conclusions to these decisions and can also identify the decisions for which it is important to determine methods a priori in future studies.

Multiverse analyses have been used across several fields of research and can address any number of potential data decisions. For example, Dejonckheere et al. (2018) conducted three studies of certain individual differences predicting symptoms of depression and used multiverse analyses to show that each of these studies found the same conclusions regardless of which scale items were used to operationalize the independent variables. In another study, Credé and Phillips (2017)

Corresponding Author:
Jenna A. Harder, Department of Psychology, Michigan State University, 316 Physics Rd., Lansing, MI 48824
E-mail: harderj3@msu.edu



Metric 1	*p* < .05	n.s.	n.s.	n.s.	*p* < .05	n.s.
Metric 2		n.s.	n.s.	*p* < .05	*p* < .05	*p* < .05
Metric 3	*p* < .05	*p* < .05	*p* < .05	n.s.	n.s.	n.s.
Team Size:	3	5	8	12	12	14

Fig. 1. Output from a hypothetical multiverse of 17 studies. Unshaded cells represent studies that supported the alternative hypothesis. Shaded cells containing "n.s." represent studies with null results. One square is left unlabeled because no study used Metric 2 with a team size of three. Results suggest that Metric 2 may be sensitive to effects in larger teams and Metric 3 may be sensitive to effects in smaller teams, whereas Metric 1 may not be a consistently effective measure.

used a multiverse analysis to test whether conclusions from a study on power poses were sensitive to rules for identifying outliers, the inclusion of control variables, and the way the dependent variable was specified. This led to the insight that conclusions about the effect of power poses on hormone levels were highly dependent on certain analytic decisions. In short, multiverse analyses are useful whenever the decisions for how best to clean and/or analyze data have an arbitrary component that could affect results.

A Multiverse of Methods

One limitation of the multiverse analysis, however, stems from the fact that only one raw data set is used to generate the multiverse. For this reason, a multiverse analysis can address only data cleaning and analytic decisions, yet researcher decisions that affect results can also happen at the data-collection stage. For example, a researcher might choose among multiple validated measures of a disorder or whether to use rats or mice as subjects. In some cases, this decision can lead to a situation in which different researchers' studies on the same question yield inconsistent results, and it is unclear which data-collection decision—if any—is responsible for the inconsistencies. The confusion may be exacerbated by additional variation across studies in other manipulated variables or in the type of statistical analysis used.

Such confusion can stall research progress by creating a lack of clarity on best practices for data collection. When there is precedent for multiple alternative methods, researchers may unwittingly choose less effective methods. Researchers may also attempt to build on the results of previous studies that used these methods, which may be methodological artifacts or entirely spurious. It would be helpful to have a tool for identifying which methodological changes actually generate different results given the same analysis. The classic multiverse analysis cannot answer this question. However, a straightforward adaptation to the multiverse approach may provide a solution.

Researchers can adapt the multiverse method to map out the impact of data-collection decisions, so that instead of multiple versions of a single data set, the multiverse of data sets is composed of real data sets from studies varying in data-collection methods of interest. The traditional multiverse of data sets is essentially replaced with what might be termed a "multiverse of methods." Subsequently applying the analysis of interest to each of these data sets would then reveal which of these decisions were consequential for study results.

For example, suppose that a field of research exists studying the effect of mindfulness training on team efficiency and that past studies have varied in (a) the size of the teams studied and (b) the metric of team efficiency. Imagine further that metrics of efficiency fall into three broad types. A researcher applying a multiverse-of-methods approach might visually represent past studies' conclusions arrayed on a grid, with three rows representing metrics of efficiency and studies arranged along those rows in order of increasing team size (see Fig. 1). This would allow the researcher to visualize how (if at all) study conclusions change with increasing team size for each of the three metrics. This multiverse analysis would indicate whether team size and efficiency metrics affect results—with implications for whether studies varying in these factors can be compared. It would also suggest hypotheses for what metrics should be used when teams of a particular size are studied.

Of course, in this simple example, similar information might be gained by performing a moderated meta-analysis. However, consider now a situation in which studies vary by five different methodological decisions, several of which are expected to interact with each other to influence results. Unless many very large data sets are available, a meta-analysis is unlikely to have enough power to quantitatively examine interactions of such complexity. This is particularly true if the data sets of interest have some hierarchical structure that must be accounted for with multilevel modeling on top of the multilevel modeling already inherent in meta-analysis. Moreover, in some situations, one question of interest is the decision about what analysis to use, which is not something that can be varied in a meta-analysis. There may even be a question of whether the outcome of different analysis options may depend on certain data-collection decisions of interest. In these situations, a multiverse analysis provides the flexibility to answer questions that a meta-analysis cannot.

The current article aims to provide a nuanced introduction to this multiverse-of-methods approach. To demonstrate the usefulness, methodology, and limitations of the approach for realistically complex areas of research, I present an extended example applying the approach to studies on shooting decisions. I begin by discussing methodological ambiguities in this line of research and how a multiverse-of-methods analysis could be applied. I then walk through an analysis applying the approach to real data from 19 shooting-decision studies. The article concludes with further discussion of some of the limitations and applications of this method.

Shooting Bias

Psychologists (Correll, Park, Judd, & Wittenbrink, 2002) have developed the First-Person Shooter Task (FPST) to measure racial bias in the kind of decisions police officers make about whether to shoot a suspect. In this laboratory task, participants view images of Black and White men holding guns or harmless objects and must quickly decide whether to press a “shoot” button if the man is holding a gun or a “don’t shoot” button if he is not. As Mekawi and Bresin (2015) confirmed in their meta-analysis, this paradigm reliably produces a pattern of responding often referred to as “shooter bias” (i.e., the tendency to choose “shoot” faster and more often for Black targets than for White targets). The past two decades have seen the growth of a subfield of researchers exploring the sources and moderators of this bias. However, despite researchers’ shared use of the same basic experimental task, the field is not unified by a common understanding of best practice in applying this task, and shooter studies vary widely in a number of both analytic and methodological practices.

Variations in analytic practice

Four analytic decisions stand out as having potential implications for the results of shooter studies. First, there is no consensus across shooter studies as to which dependent variable should be used to measure bias at the behavioral level. A shooter task produces data for two dependent variables: errors (i.e., when did the participant shoot unarmed targets and fail to shoot armed targets) and reaction time (how quickly did the participant respond). Some studies (e.g., Miller, Zielaskowski, & Plant, 2012) report analyzing only the error data. Others (e.g., Correll, Urland, & Ito, 2006) report analyzing only reaction-time data for correct responses.[1] It is not clear that one of the two response variables is the “best” variable for assessing shooter bias. However, if the choice of the response variable can affect the study’s outcome, then researchers should take care to avoid exploiting these researcher degrees of freedom.

Moreover, it is possible that reaction-time data and accuracy data may reflect bias under different circumstances. For example, studies vary considerably in the response window: the time limit for responding on a given trial. It may be that shooter bias appears in one or the other metric depending on the length of the response window. Short response windows may produce such uniformly quick responses that shooter bias appears only in errors, and long response windows may produce so few errors that shooter bias appears only in reaction time.

A second area in which shooter studies vary is the statistical methods used to analyze error data. One method is to calculate each participant’s overall error rate for each target type (armed Black, unarmed Black, armed White, unarmed White) and analyze these values within an analysis-of-variance (ANOVA) framework. A second method is to model trial-level error data, clustered by participant, with a multilevel logistical regression.[2] Modeling trial-level data with multilevel logistic regression is more appropriate statistically because summary values fail to take into account any differences across participants in the number of observations or the reliability of those observations within each condition (Nezlek, 2008). However, error-rate ANOVAs are used more often in shooter research.

The third issue again involves multilevel modeling and merits a more detailed explanation. It may be helpful to briefly review some terminology used in multilevel modeling before describing this issue. *Fixed effects* refer to traditional regression effects: intercepts and slopes for which we wish to estimate a coefficient. In multilevel modeling, any of these fixed effects can be allowed to randomly vary according to a grouping factor. For example, participant is a grouping factor if there are multiple observations for every participant. If there are multiple observations across each race-by-object combination for each participant, then it is possible to calculate an individual's personal slopes for race and for object. If we allow slopes for race to vary randomly by participant (i.e., random slopes by participant), that means we are allowing each participant to have a different slope for race. The model therefore estimates the distribution of all of these participants' race slopes and calculates the variance of all of those slope estimates. Race is the random slope, participant is the grouping factor, and the variance of those race slopes across all participants is a parameter in the model called a *random effect*. Thus, we are using participant idiosyncrasy to explain some variance in the data. Likewise, if there are multiple observations for each target—which is true in a typical shooter study because each participant sees each target multiple times—then target is a grouping factor that can also explain some variance in the data.[3]

The third issue with standard practice in shooter-study analyses is that they typically did not control for random effects of the target individual appearing in each trial. That is, the analyses did not statistically account for random variation across the stimuli themselves, which could explain differences in responses to White versus Black targets. Controlling for the variation introduced by target idiosyncrasies is important for ensuring that any observed effects can be generalized to other possible stimuli (Judd, Westfall, & Kenny, 2012)—or, in this case, to other individuals in the "real world" (i.e., suspects in police work). There is a risk that any effects found without accounting for this source of variance may be spurious, arising from some chance feature of the study's stimulus set.

This issue is further complicated by the fact that there are multiple options for how to specify random effects for targets. One option is to allow only intercepts to vary by target, such that the model allows for the possibility that each target elicits a different baseline error rate or baseline response speed (depending on which response variable is being modeled). However, the researcher can also choose to allow slopes for the object variable (i.e., whether the target is holding a gun vs. a harmless object) to vary randomly by target. Allowing object slopes to vary by target allows for the possibility that the effect of the object on the dependent variable (error or reaction time) is different for each target, even after the effect of race is accounted for. The relationship between object and error indicates the overall tendency to shoot in terms of decisions (e.g., more errors for unarmed targets means a greater frequency of the "shoot" decision), whereas the relationship between object and reaction time indicates the overall tendency to shoot in terms of speed (e.g., faster reaction times for armed targets means that "shoot" decisions are made faster than "don't shoot" decisions). Therefore, random object slopes at the target level allow for the possibility that participants' shooting responses vary across targets.

The rationale for allowing slopes for object to vary randomly by target is that a difference in shooting behavior between the Black and White conditions may arise from chance characteristics of the targets in each race. For example, suppose that several of the White targets in a shooter stimulus set happen to be less muscular, or perhaps more "baby-faced," than the average Black targets in the stimulus set. This could produce a spurious difference in how "threatening" the average member of each group appears, which could lead White targets to be shot less often (i.e., have different object slopes predicting error) than Black targets because of this aspect of appearance. However, parsing out the idiosyncratic effects of individual targets on the object slope will decrease the influence of such anomalous targets on estimates of the object slope for each race. Nevertheless, the vast majority of shooter studies have not specified random effects aside from using a within-subjects structure (i.e., allowing intercepts to vary randomly by participant).

Random-effects specifications for shooter data can also include random slopes at the participant level, and the infrequency with which this is done is a fourth issue with current analytic practice in this field. Specifically, if slopes for object are allowed to vary randomly by participant, this controls for the individual-specific relationship between object and error or speed (depending on the response variable), which can be conceptualized as individuals' baseline tendency toward a "shoot" decision. If slopes for target race are allowed to vary randomly by participant, this controls for individuals' baseline tendency toward error or speed when the target is Black versus White. Specifying analyses in this way makes it more likely that results will generalize to other participants outside the study's sample (i.e., that conclusions will not be unduly affected by oddly behaving participants). It should be acknowledged that in shooter studies, the interaction of object and race is the term of interest, and specifying random slopes for object and/or race will primarily affect the fixed main effects of those variables, with a small impact on their interaction. However, even if the impact on the interaction is negligible, researchers often report all terms of a model (e.g., in a table), and ensuring that all terms are accurately specified is advisable in the interest of correctly reporting effects that other researchers may wish to interpret.

Variation in methodological practice

Shooter studies also vary in some methodological details regarding data collection. I consider three here: response window length, number of stimuli, and sample size.

Shooter studies vary considerably in the length of the window during which participants may respond to a target, with some studies setting the response window as short as 590 ms (Ito et al., 2015) and others as long as 2,000 ms (Eiler, 2017). A few studies have even placed no limit at all on the response time (Park & Glaser, 2011; Park, Glaser, & Knowles, 2008; Park & Kim, 2015). It is unclear whether this might affect shooter bias, except for the possibility discussed above that the response window length will determine whether bias appears in reaction-time data or in accuracy data.

Another source of methodological variation is the number of stimuli used in the task. The number of different target individuals in a given shooter study can vary widely, from two (Eiler, 2017) to 50 (Sim, Correll, & Sadler, 2013). The number of targets has the potential to affect results because, as discussed in the previous section, when fewer targets are used there is a greater risk that chance variation among targets could generate spurious race effects. Moreover, even when researchers attempt to address this concern by controlling for target-level random effects, small numbers of targets pose a problem because the number of targets constrains the statistical power of these analyses. Simulations by Judd et al. (2012) suggest that shooter studies may need at least 50 targets (25 per race) to ensure 80% to 90% statistical power. Unfortunately, to my knowledge only three publications (Correll et al., 2002; Ma & Correll, 2011; Sim et al., 2013) have used this number of targets, and none have exceeded it. Likewise, participant sample sizes in shooter studies range widely, from 38 (Pleskac, Cesario, & Johnson, 2018) to 406 (Ito et al., 2015), demonstrating a lack of consensus as to the number of participants needed to detect a shooter-bias effect.

In sum, although shooter studies all purport to study the same phenomenon, there is no established norm of best practice determining how the FPST is implemented or how data from this task are analyzed. It would be useful to have some demonstration of how, if at all, study conclusions vary depending on these research decisions in real data sets. A multiverse-of-methods analysis may be helpful to this end.

Method

Overview of the multiverse

The goal of the current example was to explore both (a) the multiverse of analyses and (b) the multiverse of data-collection methods as described above. Specifically, my aim was to apply each possible analysis to as many preexisting data sets as possible. The variety of analyses would allow for an exploration of the analytic multiverse, whereas the methodological multiverse could be explored through data sets coming from studies that varied in the methodological details of interest.

Nineteen data sets were collected from various researchers. Data sets varied in participant sample size (range: $N = 38$–300), number of unique target individuals (4–50), and response window (630 ms to infinity, i.e., no enforced window). Each of these data sets was subjected to 25 different analyses. The first of these analyses was a linear regression predicting error rates (statistically equivalent to an error-rate ANOVA but performed in a regression framework to allow for the computation of regression coefficients). The other analyses included 12 different (single-level or multilevel) logistic regressions modeling error that varied in random-effects structure and 12 different (single-level or multilevel) linear regressions modeling reaction time that likewise varied in random-effects structure. Data and R code for this project are publicly available at https://osf.io/6kqxn/.

Literature search

Before gathering data sets, I referenced a meta-analysis by Mekawi and Bresin (2015) to compile a list of shooter studies for potential inclusion with publication dates up to 2012. I then conducted a literature search in PsycINFO using the quoted search terms "shooter bias," "shooting bias," "shooter task," and "shooting task" and searching for studies published after 2012. Studies were considered appropriate for inclusion if they used an FPST as described above, included both Black and White targets, manipulated target race within subjects, and had at least two targets of each race, each of whom appeared with both guns and harmless objects. Thirty-six studies were identified as potentially eligible. Of these studies, it would be possible to apply the current set of analyses only to those that had recorded which target had appeared in each trial. Authors for each of these studies were contacted to ask whether they had recorded this information and, if so, whether they would be willing to share it. Authors' responses to requests and follow-ups, combined with some data already available (i.e., data collected in my own lab), resulted in a total of 19 data sets from eight different first authors (listed in Table 1).

Multiverse analysis

A program was written in the R software environment (Version 3.5.1; R Core Team, 2018) to run each data set

Table 1. Data Sets Included in the Multiverse Analysis

Abbreviation	Study	Other manipulations	Sample size (*N*)	Targets	Length of response window (ms)
Correll1	Unpublished data—J. Correll		56	50	850
Correll2	Unpublished data—J. Correll		92	20	850
Correll11	Correll, Wittenbrink, Park, Judd, & Goyle (2011)	Background scenes were “dangerous” or “neutral”	58	20	630
HarderP	Unpublished Master’s thesis data—J. A. Harder (pilot study)	Targets varied in apparent social class	103	40	650
Harder1	Manuscript in preparation—J. A. Harder	Targets varied in apparent social class	200	40	650
Harder2	Manuscript in preparation—J. A. Harder	Targets varied in apparent social class	211	40	650
Harder3	Manuscript in preparation—J. A. Harder	Targets varied in apparent social class	101	40	650
Harder4	Manuscript in preparation—J. A. Harder	Targets varied in apparent social class	153	40	650
Kenw.1	Unpublished data—J. Kenworthy	Targets included Latino men (those trials excluded in this analysis)	96	4	850
Kenw.2	Unpublished data—J. Kenworthy	Targets included Latino men (those trials excluded in this analysis)	57	4	700
Ma1	Unpublished data—D. Ma		56	50	700
Park08	Park, Glaser, & Knowles (2008)		58	20	None
Park11	Park & Glaser (2011)	Manipulated relative frequency of counterstereotypical vs. stereotypical targets	63	20	None
Park15	Park & Kim (2015)	Manipulated whether participants were playing the video game as a White vs. Black police officer	152	20	None
Ples.17a	Pleskac, Cesario, & Johnson (2018)–Study Three	Varied discriminability of stimuli	38	46	750
Ples.17b	Pleskac, Cesario, & Johnson (2018)—Study Four	Background scenes were “dangerous” or “neutral”	108	20	630
Sim13a	Sim, Correll, & Sadler (2013)—Experiment 1	Participants read article about a White or Black person committing a violent crime	150	50	630
Sim13b	Sim, Correll, & Sadler (2013)—Experiment 2b	Manipulated relative frequency of counterstereotypical vs. stereotypical targets	122	50	630
Snow.17	Dissertation—A. Snowden	Manipulated participant emotion	300	32	730

Note: Data sets of indeterminate origin are described as unpublished data. The multiverse analysis collapsed across levels of the variables listed under the “Other manipulations” column. The number of targets shown represent the number of unique Black and White targets included in study’s stimulus set.

through each of the 25 analyses and compile the results. The object (gun/harmless object) and target race were effects-coded, and reaction time was log-transformed. The analyses used in the multiverse are listed in Table 2. The fixed effects for all analyses included the main effects for object and race as well as their interaction. The analyses used here are not exhaustive of the possible random-effects specifications;[4] however, some possible models were not used because models with more complex random-effects structures are unlikely to converge given the typical statistical power of shooter studies.

Accepted practice in work with multiverse analyses (Steegen et al., 2016) is to examine the quantitative multiverse output qualitatively. That is, the researcher charts how statistical significance varies across analyses or methods and observes the patterns that emerge. The current work follows this practice.

In addition to this, certain analyses were repeated in IBM SPSS (Version 25)[5] to obtain *p* values for the

Table 2. Random-Effects Structures Included in the Multiverse Analysis

Abbreviation	Random effects
ANOVA	None: ANOVA of participants' mean error rates for each object/race combination
Fixed	None
IP	Random intercepts by participant
IOP	Random intercepts by participant and random slopes for object by participant
IRP	Random intercepts by participant and random slopes for race by participant
IT	Random intercepts by target
IOT	Random intercepts by target and random slopes for object by target
IP + IT	Random intercepts by participant and random intercepts by target
IP + IOT	Random intercepts by participant, random intercepts by target, and random slopes for object by target
IOP + IT	Random intercepts by participant, random slopes for object by participant, and random intercepts by target
IRP + IT	Random intercepts by participant, random slopes for race by participant, and random intercepts by target
IOP + IOT	Random intercepts by participant, random slopes for object by participant, random intercepts by target, and random slopes for object by target

Note: The dependent variable was errors/reaction times for all structures except analysis of variance (ANOVA), for which the dependent variable was the mean error rate.

target-level random slopes for object. This was done for those multilevel regressions in which both intercepts and object slopes varied randomly by both target and participant (excepting those studies that experienced convergence problems with this analysis). Checking the significance of these terms was a post hoc addition to the original analysis plan but was intended to follow up one of the patterns observed in the multiverse output. If this random slope was significant for a given study, it would indicate that there was significant variation across targets in the effect object had on shooting errors.

Shooter-Bias Results and Discussion

Overview

Error-data analysis. Among analyses of error data, 218 analyses converged successfully, and 96 of these (44.0%) showed a significant Race × Object interaction, representing bias toward shooting Black targets (Fig. 2). Among the 96 analyses that found a significant interaction, the mean regression coefficient was −0.107 ($SD = 0.066$; $e^{-0.10} = 0.899$), which would correspond to a decrease of about 10% in the odds of making an error when race and object are stereotype-congruent (i.e., a Black person holding a gun or a White person holding a harmless object).

Reaction-time analyses. Among analyses of reaction-time data, 224 analyses converged successfully, 73 of which (32.6%) showed a significant Race × Object interaction (Fig. 3). Of these, 11 (across two studies) indicated a bias toward shooting White targets, and 62 indicated a bias toward shooting Black targets. The mean interaction coefficient of the 62 analyses showing bias toward shooting Black targets was −0.010 ($SD = 0.005$),[6] indicating that responses were faster by 10 ms when targets were stereotype-congruent. The mean interaction coefficient of the 11 studies showing bias toward shooting White targets was 0.013 ($SD = 0.004$).

Because relatively few analyses indicated a bias toward shooting White targets, and because these analyses came from only two studies that were not similar in any of the measured variables, it is not possible from these data to identify predictors of the direction of the coefficient. Therefore, the discussion below focuses on predictors of statistically significant Race × Object interaction coefficients, including all analyses without regard to the direction of the effect.

Type of analysis and number of targets

Participant as a grouping factor. Although participant-level random effects (see Table 3) were generally significant, the precise specification of effects at this level did not seem to be related to the significance of the Race × Object interaction in any way.[7] That is, among logistic regressions predicting error with no target-level random effects, interactions were significant in 11 (61.1%) of 18 converging analyses with random intercepts by participant (Fig. 2); and of 17 converging analyses with random intercepts and object slopes by participant or with random intercepts and race slopes by participant, interactions were significant in 10 studies (58.8%). Likewise, among 19 linear-regression analyses predicting reaction time, analyses with random intercepts by participant or with random intercepts and object slopes by participant yielded significant results in nine studies (47.4%), and analyses with random intercepts and race slopes by participants yielded significant results in eight studies (42.1%; Fig. 3). This overall lack of a pattern is unsurprising, as

	← Shorter Windows									Study							Longer Windows →		
Error Analysis	Correll11	Ples.17b	Sim13a	Sim13b	Harder2	Harder3	HarderP	Harder1	Harder4	Kenw.1	Kenw.2	Snow.17	Ples.17a	Correll1	Correll2	Ma1	Park08	Park11	Park15
ANOVA	.061	.023	< .001	< .001	.326	.160	.720	.607	.710	.063	.016	.213	.325	.094	.005	.160	.704	.398	.764
Fixed	.001	< .001	< .001	< .001	.023	< .001	.508	.122	.927	.017	.001	.038	.016	.076	< .001	.122	.354	.327	.403
IP	< .001	< .001	< .001	< .001	.020	< .001	.502	.118	.894	.016	.001	.028	.014	.069	< .001	.116		.336	.401
IOP	< .001	< .001	< .001	< .001	.015	< .001	.466	.129	.868	.016	.001	.032	.013	.083		.119	.394	.335	
IRP	< .001	< .001		< .001	.027	< .001	.514	.095		.013	.001	.028	.014	.085	< .001	.121	.324	.336	.400
IT	.001	< .001	< .001	< .001	.016	< .001	.159	.158	.810	.029	.001	.037	.015	.052	< .001	.093	.355	.342	.385
IP+IT	< .001	< .001	< .001	< .001	.014	< .001	.153	.153	.776	.029	.002	.027	.012		< .001	.083		.352	.381
IOP+IT	< .001	< .001	< .001	< .001	.010	< .001	.133	.168	.743	.028	.002	.030	.011	.091	< .001	.109		.348	.379
IRP+IT	< .001		< .001	< .001	.019	< .001	.158	.127	.873	.025	.002	.027	.011	.063	< .001		.326		.381
IOT	.118	.259	.076	.094	.490	.280	.892	.796	.890	.444	.003	.306	.274	.774	.011				
IP+IOT	.111	.257	.075	.068	.491	.271	.892	.796	.889	.449	.004	.307	.265	.766	.005		.464	.222	.500
IOP+IOT	.118	.257	.078	.064	.485	.278	.909	.807	.881	.446	.004	.317	.263	.780		.640			
IRP+IOT	.109	.254	.070	.066	.509	.276	.889	.779			.004		.265						
Window:	630	630	630	630	650	650	650	650	650	700	700	730	750	850	850	850	None	None	None
Sample Size:	58	108	150	122	211	101	103	200	153	96	57	300	38	56	92	56	58	63	152
Number Targets:	10	10	25	25	20	20	20	20	20	2	2	16	23	25	10	25	10	10	10

Fig. 2. Results from 13 analyses predicting errors in 19 studies, ordered by sample size. Values are *p* values from the Race × Object interaction representing racial bias in shooting decisions. Each cell represents a single analysis/study combination. See Table 1 for descriptions of the abbreviated study citations. See Table 2 for descriptions of the abbreviated analyses. Cells representing nonsignificant results are shaded light gray; cells representing significant bias toward shooting Black targets are unshaded. Blank cells indicate analyses that did not converge.

Shorter Windows ← Study → Longer Windows

Reaction Time Analysis	Correll11	Ples.17b	Sim13a	Sim13b	Harder2	Harder3	HarderP	Harder1	Harder4	Kenw.1	Kenw.2	Snow.17	Ples.17a	Correll1	Correll2	Ma1	Park08	Park11	Park15
Fixed	.426	.118	< .001	.007	.472	.011	.080	.548	.517	< .001	< .001	.114	.990	< .001	< .001	< .001	< .001	.217	.382
IP	.452	.096	< .001	.001	.730	.105	.025	.535	.606	< .001	< .001	.083	.819	< .001	< .001	< .001	< .001	.655	.400
IOP	.408	.088	< .001	< .001	.640	.154	.027	.419	.636	< .001	< .001	.074	.815	< .001	< .001	< .001	< .001	.685	.401
IRP	.441	.097	< .001	.001	.821	.084	.109	.521	.583	< .001	< .001	.083	.818	< .001	< .001	< .001	< .001	.671	.402
IT	.398	.156	< .001	.005	.488	.011	.123	.648	.507	< .001	.001	.115	.583	< .001	< .001	< .001	< .001	.185	.365
IP+IT	.424	.154	< .001	< .001	.760	.102	.088	.638	.590	< .001	< .001	.087	.407	< .001	< .001	< .001	< .001	.516	.380
IOP+IT	.378	.143	< .001	< .001	.668	.146	.102	.510	.615	< .001	< .001	.077	.399	< .001	< .001	< .001	< .001	.542	.381
IRP+IT	.411	.155	< .001	.001	.852	.083	.275	.623	.569	< .001	< .001	.087	.408	< .001	< .001	< .001	< .001	.532	.381
IOT	.599	.386	.241	.162		.168		.934	.740	.215	.095	.459	.924	.042	.219	.049	.089		.825
IP+IOT	.618	.478	.215	.114	.937	.481	.168	.971	.779	.224	.093	.475	.851	.043	.205	.051	.089	.821	.833
IOP+IOT	.604	.481	.206	.103	.885	.531	.182	.916	.790	.225	.096	.467	.860	.043	.205	.050	.094	.841	.833
IRP+IOT	.610	.481	.214	.129	.986	.452	.340	.965	.773	.224		.476	.852	.043	.205	.051	.092	.828	.834
Window:	630	630	630	630	650	650	650	650	650	700	700	730	750	850	850	850	None	None	None
Sample Size:	58	108	150	122	211	101	103	200	153	96	57	300	38	56	92	56	58	63	152
Number Targets:	10	10	25	25	20	20	20	20	20	2	2	16	23	25	10	25	10	10	10

Fig. 3. Results from 12 analyses predicting reaction times in 19 studies, ordered by response window. Values indicate *p* values from the Race × Object interaction representing racial bias in shooting-response latencies. Each cell represents a single analysis/study combination. See Table 1 for descriptions of the abbreviated study citations. See Table 2 for descriptions of the abbreviated analyses. Cells representing nonsignificant results are shaded light gray; cells representing significant bias toward shooting Black targets are unshaded. Cells representing significant bias toward shooting White targets are shaded dark gray. Blank cells indicate analyses that did not converge.

Table 3. Random Effects for Models With Object Slopes Varying by Both Target and Participant

	Random intercepts		Random slopes	
Study	By participant	By target	For object by participant	For object by target
		Errors		
Correll1	.725 (.207)***	.518 (.155)***	.541 (.172)**	.078 (.062)
Correll11	.090 (.037)*	.326 (.074)***	.088 (.036)*	.120 (.034)***
HarderP	.112 (.029)***	.130 (.022)***	.089 (.024)***	.238 (.037)***
Harder1	.096 (.023)***	.170 (.019)***	.068 (.016)***	.159 (.018)***
Harder2	.088 (.021)***	.017 (.018)***	.046 (.011)***	.151 (.016)***
Harder3	.123 (.031)***	.522 (.080)***	.124 (.032)***	.181 (.031)***
Harder4	.181 (.043)***	.439 (.053)***	.151 (.036)***	.042 (.007)***
Ma1	.466 (.131)***	.357 (.104)***	.434 (.129)***	.081 (.048)
Kenw.1	.028 (.029)	.226 (.055)***	.081 (.075)	.029 (.024)
Ples.17a	.246 (.065)***	.701 (.177)***	.156 (.045)***	.028 (.017)
Ples.17b	.059 (.021)**	.228 (.034)***	.068 (.024)**	.105 (.017)***
Sim13a	.307 (.065)***	.336 (.044)***	.388 (.082)***	.078 (.013)***
Sim13b	.215 (.045)***	.439 (.060)***	.274 (.057)***	.128 (.020)***
Snow.17	.068 (.019)***	.540 (.051)***	.030 (.010)**	.072 (.012)***
		Response times		
Correll1	.003 (.001)***	.003 (.001)***	< .001 (< .001)***	.003 (.001)***
Correll2	.003 (.001)***	< .001 (< .001)**	< .001 (< .001)***	.001 (< .001)**
Correll11	.001 (< .001)***	< .001 (< .001)**	< .001 (< .001)***	< .001 (< .001)**
HarderP	.030 (.004)***	< .001 (< .001)*	.004 (.001)***	< .001 (< .001)*
Harder1	.010 (.001)***	< .001 (< .001)***	.001 (< .001)***	< .001 (< .001)***
Harder2	.019 (.002)***	< .001 (< .001)**	.002 (< .001)***	< .001 (< .001)***
Harder3	.030 (.005)***	< .001 (< .001)**	.002 (.001)***	< .001 (< .001)***
Harder4	.003 (< .001)***	< .001 (< .001)***	.001 (< .001)***	< .001 (< .001)***
Kenw.1	.002 (< .001)***	< .001 (< .001)	< .001 (< .001)**	< .001 (< .001)
Kenw.2	.001 (< .001)***	< .001 (< .001)	< .001 (< .001)*	< .001 (< .001)
Ma1	.003 (.001)***	.003 (.001)***	< .001 (< .001)***	.003 (.001)***
Park08	.021 (.003)***	.001 (< .001)*	.004 (.001)***	.001 (.001)*
Park11	.026 (.005)***	.001 (.001)*	.005 (.001)***	.002 (.001)*
Ples.17a	.001 (< .001)***	.002 (< .001)***	< .001 (< .001)**	.001 (< .001)***
Ples.17b	.011 (.002)***	< .001 (< .001)*	.001 (< .001)***	< .001 (< .001)*
Sim13a	.002 (< .001)***	.001 (< .001)***	< .001 (< .001)***	.001 (< .001)***
Sim13b	.006 (.001)***	.001 (< .001)***	< .001 (< .001)***	.001 (< .001)***
Snow.17	.129 (.012)***	.013 (.004)***	.015 (.003)***	.011 (.003)**

Note: Values in parentheses are standard errors. See Table 1 for descriptions of the abbreviated study citations. Errors represent multilevel logistic-regression coefficients for the interaction between target race and target object predicting whether the participant made an error on a given trial. Response times represent multilevel linear-regression coefficients for the Race × Object interaction predicting the participant's response time on a given trial. Each coefficient comes from a model with one of four random-effects structures, indicated by the column headings. If random slopes were included for a grouping factor, random intercepts were also included for that grouping factor.

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

allowing either the object or the race main effect to vary by participant should primarily affect the degrees of freedom (and therefore the significance) of that main effect rather than Race × Object interaction effects. Nonetheless, this has reassuring implications for past shooter studies that did not allow slopes to vary by participant, suggesting that this analytic misspecification may not have affected the accuracy of the results.

Target as a grouping factor. On the other hand, allowing object main effects to vary by target should affect the Race × Object interaction because race is implicit in target identity. That is, when the grouping factor (target) contains information about one variable (race), allowing variance in the slope of the other variable (object) across levels of that grouping factor (i.e., allowing each target to have a different slope for object) will affect the degrees of freedom for the interaction of those two variables. Thus, for both error data and reaction-time data there was a striking difference between analyses that did versus did not control for random object slopes at the target level (Figs. 2 and 3). Across analyses that did control for random object slopes at the target level, Race × Object interactions were significant in only 11.9% of converging reaction-time analyses and 10.3% of converging-error analyses, although among analyses with simpler random-effects structures, 45.4% of reaction-time analyses and 56.3% of error analyses found significant Race × Object interactions.

As stated earlier, differences in significance in a multiverse analysis do not themselves reveal the "right" way to analyze shooter data. That is, a multiverse analysis does not speak to whether a given significant result is a true positive or a false positive or whether a given nonsignificant result is a true or false negative. Such judgments must be based on an understanding of the analyses in question. In this case, there is good reason to believe that using more complicated random-effects structures is the best way to analyze data. However, it is also true that because of the small numbers of targets that characterized nearly every study, these studies were not sufficiently powered to detect shooter bias in such analyses.

Number of targets. Perhaps relatedly, results did not appear to differ between studies with larger numbers of targets and studies with fewer targets; this was true across the various analyses and for both reaction-time data and error data (Figs. 4 and 5). For reaction-time data, 34.2% of analyses from studies at or above the median number of targets (40) showed a significant Race × Object interaction compared with 29.6% of analyses from studies below the median. For error data, the interaction was significant in 31.5% of analyses from studies above the median versus 47.0% of analyses from studies below the median. Interestingly, in the analyses of error data, studies with few targets may have yielded significant effects somewhat more often than studies with more targets. There is no clear reason why this pattern should have emerged; it may be due to the confounding influence of some unmeasured study-level variable. It is uncertain whether this observation represents a true pattern that would generalize to other data sets.

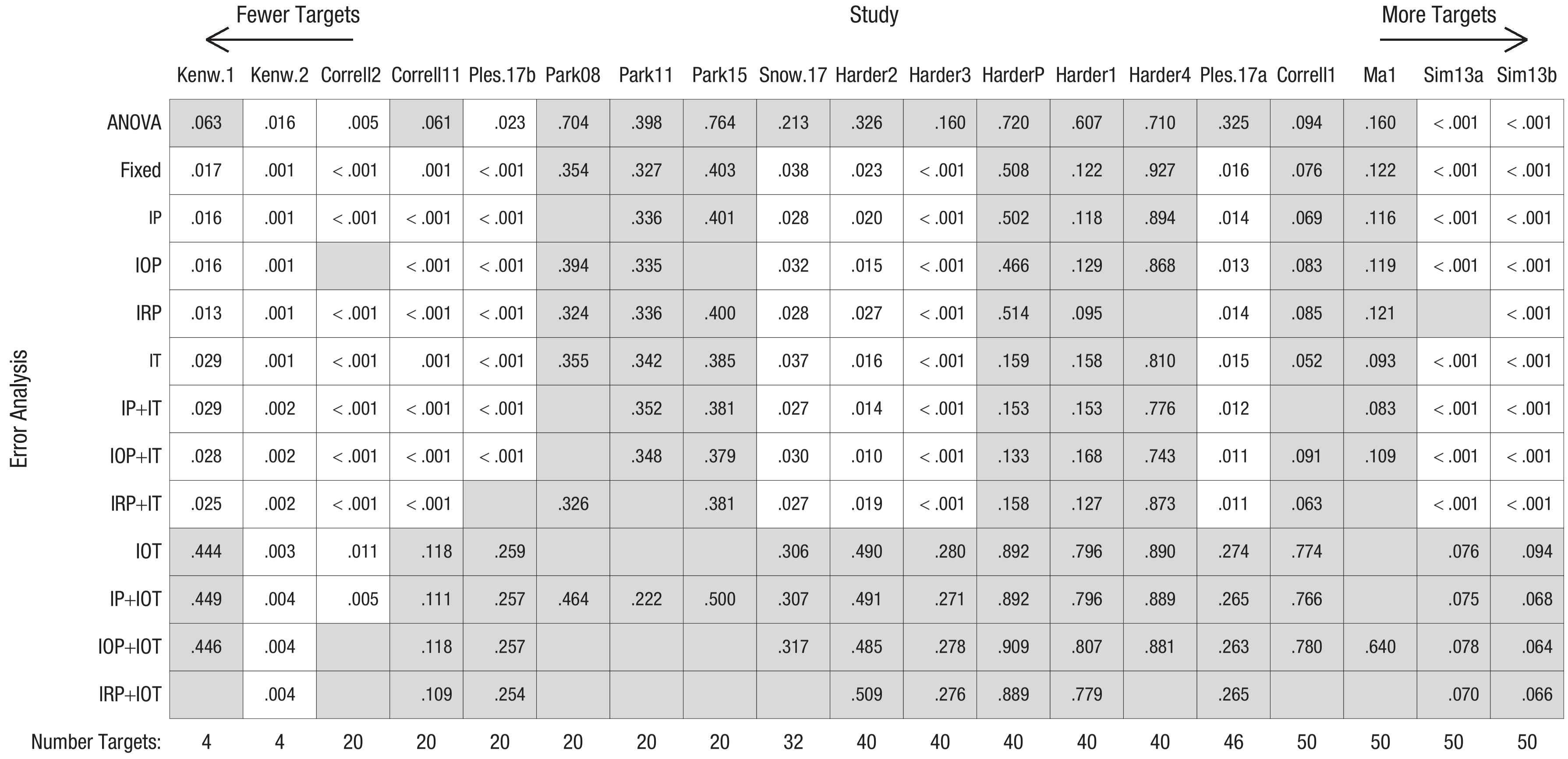
It is therefore difficult to draw conclusions about whether participants exhibited shooter bias in these studies. The analyses with simpler random-effects structures increase Type I error risks because they do not account for extraneous sources of variance, but the analyses with sufficiently complex random-effects structures increase Type II error risks because they are underpowered. This ambiguity demonstrates the importance of considering these issues when designing shooter studies. With respect to the studies analyzed here, the significant results of the regressions with simple random-effects structures should be treated with some skepticism when they are not matched with significant results from regressions with more complete random-effects structures. In other words, the fixed effects that here became nonsignificant when object slopes vary randomly by target may or may not represent "true" effects.

Type of analysis. Finally, another pattern that emerged in the multiverse of analyses was that results of error-rate ANOVAs for Race × Object interactions were found to be significant less often than most of the trial-level regressions (with the exception of those that allowed slopes to vary by target; see Fig. 2). Interactions were significant in 5 of 19 error-rate ANOVAs compared with, for example, 11 of 17 converging logistic regressions with random intercepts by participant and target. It should be noted that results from the multilevel logistic regressions are more likely to be accurate than results from the error-rate ANOVAs. Analyses of mean values, such as ANOVAs of error rates, are problematic because they do not account for differences in reliability among the means (Nezlek, 2001).

Response window

Multiverse results indicated that the response window (Figs. 2 and 3) may, as hypothesized, be important in whether race bias appears in error data versus reaction-time data. The hypothesis for error analyses was that lower response windows would be associated with a greater frequency of significant Race × Object interactions. Such a pattern was not obvious from an examination of a median split of error analyses, given that 43.4% of converging error analyses from studies below the median response window (700 ms) showed a significant Race × Object interaction compared with 43.8% from studies above the median. However, a difference does seem to exist between response windows above versus below 850 ms. Among the six studies with response windows of 850 ms or higher, only 18.2% of converging error analyses showed a significant interaction.

The hypothesis for reaction-time analyses, in contrast, was that longer response windows would be associated with a greater frequency of significant Race × Object interactions. Consistent with this hypothesis were results that showed a significant Race × Object



Fewer Targets ← Study → More Targets

Error Analysis	Kenw.1	Kenw.2	Correll2	Correll11	Ples.17b	Park08	Park11	Park15	Snow.17	Harder2	Harder3	HarderP	Harder1	Harder4	Ples.17a	Correll1	Ma1	Sim13a	Sim13b
ANOVA	.063	.016	.005	.061	.023	.704	.398	.764	.213	.326	.160	.720	.607	.710	.325	.094	.160	< .001	< .001
Fixed	.017	.001	< .001	.001	< .001	.354	.327	.403	.038	.023	< .001	.508	.122	.927	.016	.076	.122	< .001	< .001
IP	.016	.001	< .001	< .001	< .001		.336	.401	.028	.020	< .001	.502	.118	.894	.014	.069	.116	< .001	< .001
IOP	.016	.001		< .001	< .001	.394	.335		.032	.015	< .001	.466	.129	.868	.013	.083	.119	< .001	< .001
IRP	.013	.001	< .001	< .001	< .001	.324	.336	.400	.028	.027	< .001	.514	.095		.014	.085	.121		< .001
IT	.029	.001	< .001	.001	< .001	.355	.342	.385	.037	.016	< .001	.159	.158	.810	.015	.052	.093	< .001	< .001
IP+IT	.029	.002	< .001	< .001	< .001		.352	.381	.027	.014	< .001	.153	.153	.776	.012		.083	< .001	< .001
IOP+IT	.028	.002	< .001	< .001	< .001		.348	.379	.030	.010	< .001	.133	.168	.743	.011	.091	.109	< .001	< .001
IRP+IT	.025	.002	< .001	< .001		.326		.381	.027	.019	< .001	.158	.127	.873	.011	.063		< .001	< .001
IOT	.444	.003	.011	.118	.259				.306	.490	.280	.892	.796	.890	.274	.774		.076	.094
IP+IOT	.449	.004	.005	.111	.257	.464	.222	.500	.307	.491	.271	.892	.796	.889	.265	.766		.075	.068
IOP+IOT	.446	.004		.118	.257				.317	.485	.278	.909	.807	.881	.263	.780	.640	.078	.064
IRP+IOT		.004		.109	.254					.509	.276	.889	.779		.265			.070	.066
Number Targets:	4	4	20	20	20	20	20	20	32	40	40	40	40	40	46	50	50	50	50

Fig. 4. Results (*p* values) from 13 analyses predicting errors in 19 studies, ordered by number of targets. Each cell represents a single analysis/study combination. See Table 1 for descriptions of the abbreviated study citations. See Table 2 for descriptions of the abbreviated analyses. Cells representing nonsignificant results are shaded light gray; cells representing significant bias toward shooting Black targets are shaded dark gray. Blank cells indicate analyses that did not converge.

← Fewer Targets | Study | More Targets →

Reaction Time Analysis	Kenw.1	Kenw.2	Correll2	Correll11	Ples.17b	Park08	Park11	Park15	Snow.17	Harder2	Harder3	HarderP	Harder1	Harder4	Ples.17a	Correll1	Ma1	Sim13a	Sim13b
Fixed	< .001	< .001	< .001	.426	.118	< .001	.217	.382	.114	.472	.011	.080	.548	.517	.990	< .001	< .001	< .001	.007
IP	< .001	< .001	< .001	.452	.096	< .001	.655	.400	.083	.730	.105	.025	.535	.606	.819	< .001	< .001	< .001	.001
IOP	< .001	< .001	< .001	.408	.088	< .001	.685	.401	.074	.640	.154	.027	.419	.636	.815	< .001	< .001	< .001	< .001
IRP	< .001	< .001	< .001	.441	.097	< .001	.671	.402	.083	.821	.084	.109	.521	.583	.818	< .001	< .001	< .001	.001
IT	< .001	.001	< .001	.398	.156	< .001	.185	.365	.115	.488	.011	.123	.648	.507	.583	< .001	< .001	< .001	.005
IP+IT	< .001	< .001	< .001	.424	.154	< .001	.516	.380	.087	.760	.102	.088	.638	.590	.407	< .001	< .001	< .001	< .001
IOP+IT	< .001	< .001	< .001	.378	.143	< .001	.542	.381	.077	.668	.146	.102	.510	.615	.399	< .001	< .001	< .001	< .001
IRP+IT	< .001	< .001	< .001	.411	.155	< .001	.532	.381	.087	.852	.083	.275	.623	.569	.408	< .001	< .001	< .001	.001
IOT	.215	.095	.219	.599	.386	.089		.825	.459		.168		.934	.740	.924	.042	.049	.241	.162
IP+IOT	.224	.093	.205	.618	.478	.089	.821	.833	.475	.937	.481	.168	.971	.779	.851	.043	.051	.215	.114
IOP+IOT	.225	.096	.205	.604	.481	.094	.841	.833	.467	.885	.531	.182	.916	.790	.860	.043	.050	.206	.103
IRP+IOT	.224		.205	.610	.481	.092	.828	.834	.476	.986	.452	.340	.965	.773	.852	.043	.051	.214	.129
Number Targets:	4	4	20	20	20	20	20	20	32	40	40	40	40	40	46	50	50	50	50

Fig. 5. Results (*p* values) from 12 analyses predicting reaction times in 19 studies, ordered by number of targets. Each cell represents a single analysis/study combination. See Table 1 for descriptions of the abbreviated study citations. See Table 2 for descriptions of the abbreviated analyses. Cells representing nonsignificant results are shaded light gray; cells representing significant bias toward shooting Black targets are unshaded. Cells representing significant bias toward shooting White targets are shaded dark gray. Blank cells indicate analyses that did not converge.

interaction in 18.9% of converging reaction-time analyses from studies below the median compared with 44.9% above the median. In fact, among studies with response windows of 850 ms or higher, 52.1% of reaction-time analyses showed a significant interaction compared with only 18.2% of converging error analyses. Taken as a whole, the results suggest that response windows at or above 850 ms are more likely to reveal bias in response times, whereas response windows below 700 ms are more likely to reveal bias in errors (no clear pattern emerged for response windows between 700 and 850 ms). This pattern suggests that researchers should choose their response windows differently depending on their outcome of interest.

Sample size

Error-data analyses. An examination of the pattern of logistic-regression results across sample sizes (Fig. 6) indicates that studies with larger samples do not seem to produce significant Race × Object interactions in errors more often than studies with smaller samples. Among studies at or above the median sample size ($N = 101$), 40.8% of converging error analyses yielded a significant Race × Object interaction in some analyses compared with 48.5% of analyses among studies below that sample size. The lack of a sample-size effect among the error analyses may at first appear surprising, as one would expect studies with larger samples to have higher power to detect a true effect. Moreover, a range restriction was not a problem in this set of sample sizes, which ranged from 38 to 300. However, for those analyses that controlled for target-level random effects (intercepts and/or slopes), it may be at least partially explained by the fact that the power of studies to detect effects in these analyses is limited by the number of unique targets. Simulations from Judd et al. (2012) suggest that when target random intercepts are specified, the statistical power for studies with a given number of targets begins to approach an asymptote for sample sizes greater than about 30. That is, after this point, the amount of power that can be gained from increasing the number of participants is relatively small. All of the studies included here had sample sizes of 38 and above, so power may not have been strongly related to sample size for the majority of the analyses. However, this does not explain why sample size was also unrelated to significance for analyses that did not specify random effects for target. Some degree of publication bias may be to blame here.

Reaction-time analyses. In reaction-time analyses (Fig. 7), 15.6% of converging analyses from studies at or above the median sample size yielded a significant interaction compared with 50.0% from studies with smaller sample sizes. That is, studies with larger samples (and thus higher power) were actually slightly less likely to find an effect. This pattern, however, should be interpreted with caution, as there is no statistical or methodological reason to expect such an effect of high sample sizes. It may be that some unmeasured factor covaries with sample size and is playing a confounding role.

Recommendations for shooter studies

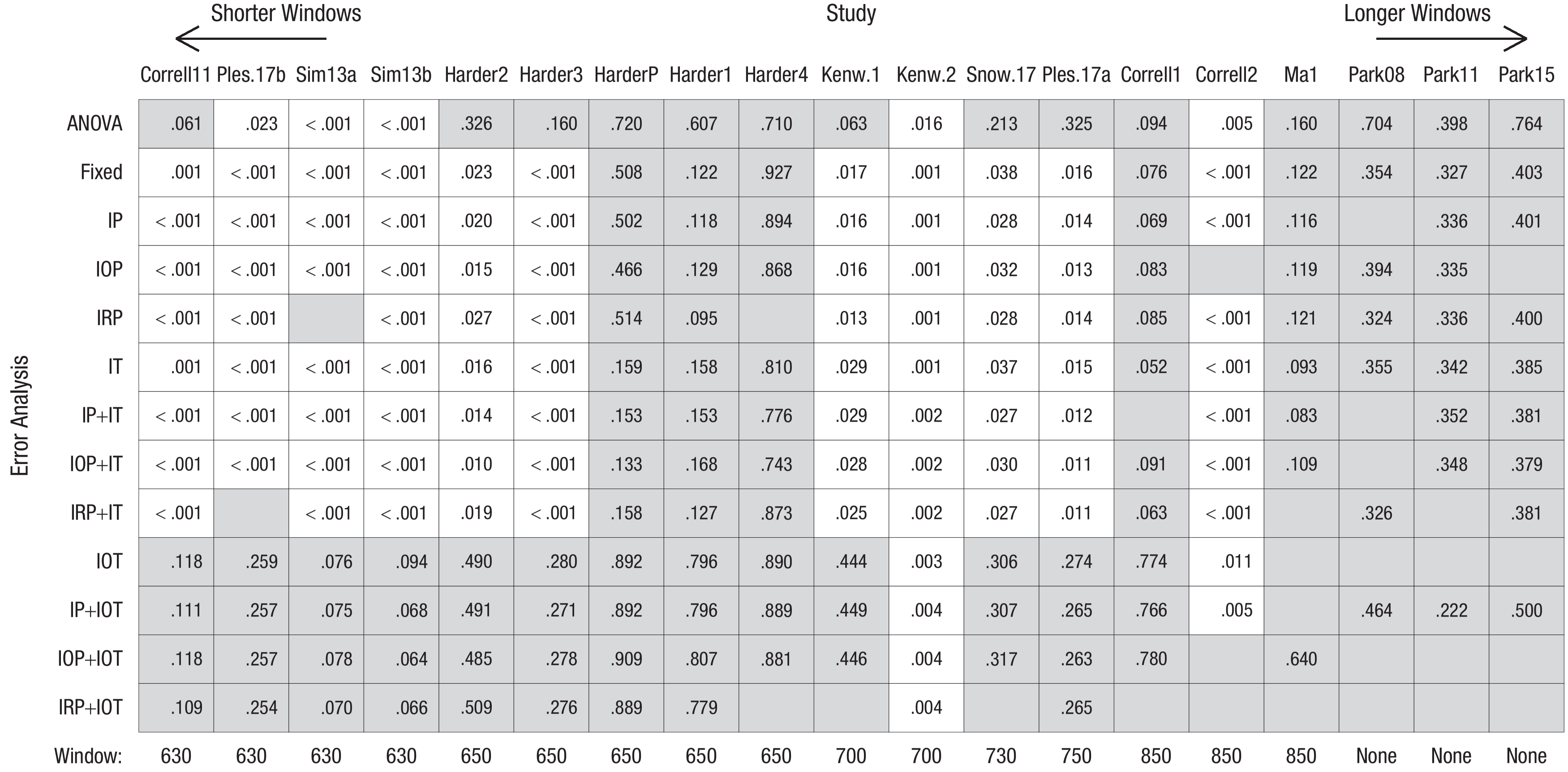
In light of the current findings and the discussion above, certain recommendations seem reasonable. First, shooter studies should use more than 50 targets. Judd et al. (2012) indicated that at least 50 targets were necessary to secure adequate power in a shooter study. The studies examined here, however, all used 50 or fewer targets, and very few showed any significant shooter bias when object slopes were allowed to vary randomly by target. If the studies had used larger numbers of targets, this problem might have been prevented.

Second, the choice of which response window to use should depend on the response variable of interest, as shooter bias was distributed differently across response window for error versus reaction-time data. To maximize the probability of detecting an effect should one exist, researchers interested in examining bias in shooting errors should use a short response window, such as 630 or 650 ms. Researchers interested in examining bias in shooting response times, however, should use a long response window, such as 850 ms.

Third, the current results suggest that shooter researchers should use multilevel regression models—as opposed to ANOVAs of summary data—when analyzing data. Moreover, random-effects structures should be as complex as the data set can handle without the models' experiencing convergence problems.[8] Nevertheless, results do indicate that specifying target-level slopes for object is more important to conclusions about racial bias than specifying participant-level effects. Thus, in situations in which random-effects structures must be simplified (e.g., if a researcher finds that the computational intensity of specifying random slopes at both the participant level and the target level is too great), it would be less likely to affect results if the researcher were to sacrifice the participant-level slopes rather than the target-level slopes.

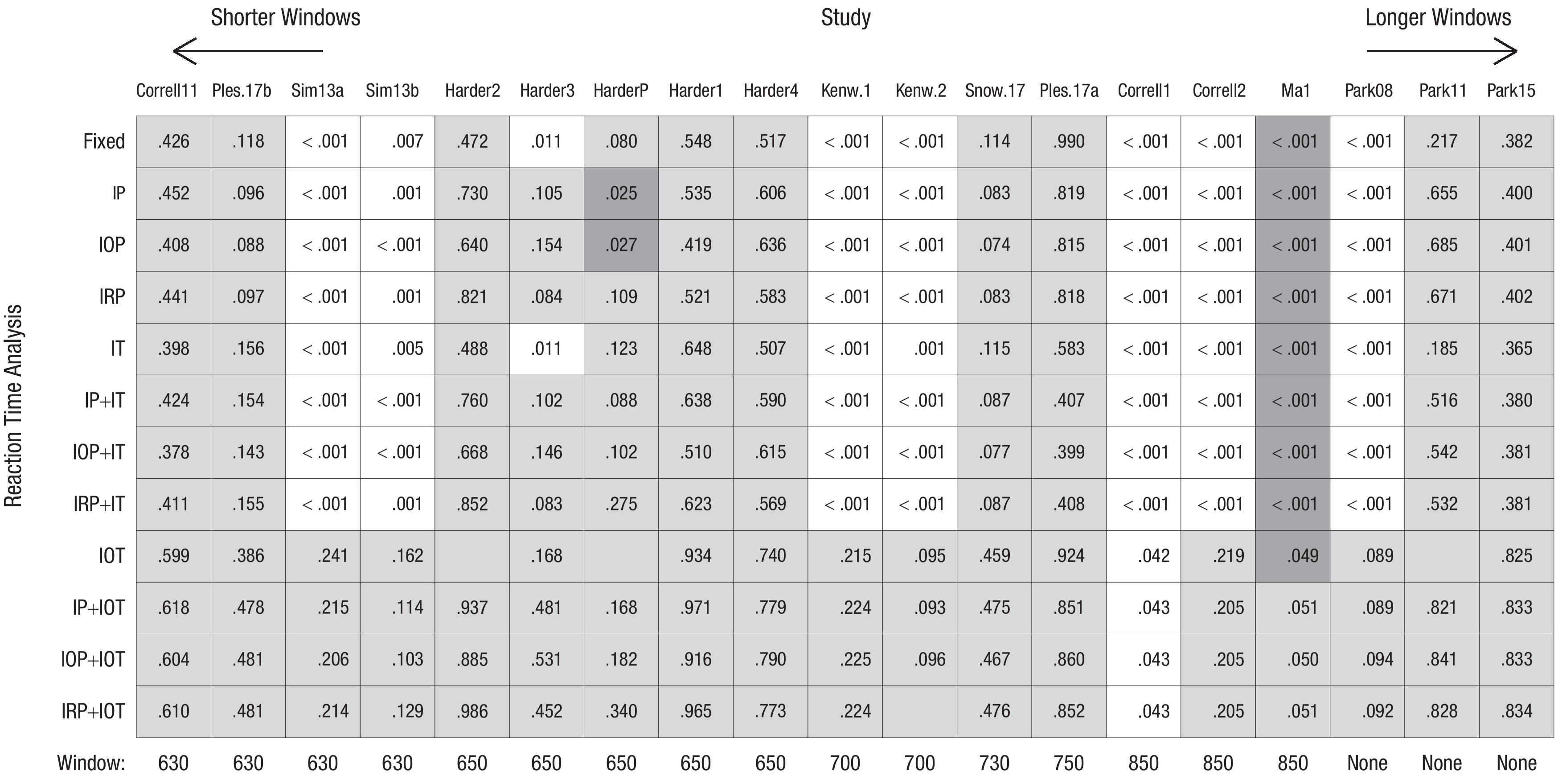
General Discussion

The shooter-bias example illustrates how examining a multiverse of methods can inform methodological decisions. It also illustrates how examining a multiverse of analyses can deepen the researcher's understanding of statistical decisions. The results of this example suggested


Shorter Windows
Study
Longer Windows
Error Analysis

	Correll11	Ples.17b	Sim13a	Sim13b	Harder2	Harder3	HarderP	Harder1	Harder4	Kenw.1	Kenw.2	Snow.17	Ples.17a	Correll1	Correll2	Ma1	Park08	Park11	Park15
ANOVA	.061	.023	< .001	< .001	.326	.160	.720	.607	.710	.063	.016	.213	.325	.094	.005	.160	.704	.398	.764
Fixed	.001	< .001	< .001	< .001	.023	< .001	.508	.122	.927	.017	.001	.038	.016	.076	< .001	.122	.354	.327	.403
IP	< .001	< .001	< .001	< .001	.020	< .001	.502	.118	.894	.016	.001	.028	.014	.069	< .001	.116		.336	.401
IOP	< .001	< .001	< .001	< .001	.015	< .001	.466	.129	.868	.016	.001	.032	.013	.083		.119	.394	.335	
IRP	< .001	< .001		< .001	.027	< .001	.514	.095		.013	.001	.028	.014	.085	< .001	.121	.324	.336	.400
IT	.001	< .001	< .001	< .001	.016	< .001	.159	.158	.810	.029	.001	.037	.015	.052	< .001	.093	.355	.342	.385
IP+IT	< .001	< .001	< .001	< .001	.014	< .001	.153	.153	.776	.029	.002	.027	.012		< .001	.083		.352	.381
IOP+IT	< .001	< .001	< .001	< .001	.010	< .001	.133	.168	.743	.028	.002	.030	.011	.091	< .001	.109		.348	.379
IRP+IT	< .001		< .001	< .001	.019	< .001	.158	.127	.873	.025	.002	.027	.011	.063	< .001		.326		.381
IOT	.118	.259	.076	.094	.490	.280	.892	.796	.890	.444	.003	.306	.274	.774	.011				
IP+IOT	.111	.257	.075	.068	.491	.271	.892	.796	.889	.449	.004	.307	.265	.766	.005		.464	.222	.500
IOP+IOT	.118	.257	.078	.064	.485	.278	.909	.807	.881	.446	.004	.317	.263	.780		.640			
IRP+IOT	.109	.254	.070	.066	.509	.276	.889	.779			.004		.265						
Window:	630	630	630	630	650	650	650	650	650	700	700	730	750	850	850	850	None	None	None

Fig. 6. Results (*p* values) from 13 analyses predicting errors in 19 studies, ordered by sample size. Each cell represents a single analysis/study combination. See Table 1 for descriptions of the abbreviated study citations. See Table 2 for descriptions of the abbreviated analyses. Cells representing nonsignificant results are shaded light gray; cells representing significant bias toward shooting Black targets are unshaded. Blank cells indicate analyses that did not converge.


Shorter Windows
Study
Longer Windows
Reaction Time Analysis

	Correll11	Ples.17b	Sim13a	Sim13b	Harder2	Harder3	HarderP	Harder1	Harder4	Kenw.1	Kenw.2	Snow.17	Ples.17a	Correll1	Correll2	Ma1	Park08	Park11	Park15
Fixed	.426	.118	< .001	.007	.472	.011	.080	.548	.517	< .001	< .001	.114	.990	< .001	< .001	< .001	< .001	.217	.382
IP	.452	.096	< .001	.001	.730	.105	.025	.535	.606	< .001	< .001	.083	.819	< .001	< .001	< .001	< .001	.655	.400
IOP	.408	.088	< .001	< .001	.640	.154	.027	.419	.636	< .001	< .001	.074	.815	< .001	< .001	< .001	< .001	.685	.401
IRP	.441	.097	< .001	.001	.821	.084	.109	.521	.583	< .001	< .001	.083	.818	< .001	< .001	< .001	< .001	.671	.402
IT	.398	.156	< .001	.005	.488	.011	.123	.648	.507	< .001	.001	.115	.583	< .001	< .001	< .001	< .001	.185	.365
IP+IT	.424	.154	< .001	< .001	.760	.102	.088	.638	.590	< .001	< .001	.087	.407	< .001	< .001	< .001	< .001	.516	.380
IOP+IT	.378	.143	< .001	< .001	.668	.146	.102	.510	.615	< .001	< .001	.077	.399	< .001	< .001	< .001	< .001	.542	.381
IRP+IT	.411	.155	< .001	.001	.852	.083	.275	.623	.569	< .001	< .001	.087	.408	< .001	< .001	< .001	< .001	.532	.381
IOT	.599	.386	.241	.162		.168		.934	.740	.215	.095	.459	.924	.042	.219	.049	.089		.825
IP+IOT	.618	.478	.215	.114	.937	.481	.168	.971	.779	.224	.093	.475	.851	.043	.205	.051	.089	.821	.833
IOP+IOT	.604	.481	.206	.103	.885	.531	.182	.916	.790	.225	.096	.467	.860	.043	.205	.050	.094	.841	.833
IRP+IOT	.610	.481	.214	.129	.986	.452	.340	.965	.773	.224		.476	.852	.043	.205	.051	.092	.828	.834
Window:	630	630	630	630	650	650	650	650	650	700	700	730	750	850	850	850	None	None	None

Fig. 7. Results (*p* values) from 12 analyses predicting reaction times in 19 studies, ordered by sample size. Each cell represents a single analysis/study combination. See Table 1 for descriptions of the abbreviated study citations. See Table 2 for descriptions of the abbreviated analyses. Cells representing nonsignificant results are shaded light gray; cells representing significant bias toward shooting Black targets are unshaded. Cells representing significant bias toward shooting White targets are shaded dark gray. Blank cells indicate analyses that did not converge.

that the effect of response window is indeed moderated by the response variable, which clarified the importance of selecting a response window with the response variable of interest in mind. Moreover, the multiverse analysis not only confirmed that decisions about how to specify random effects can affect the outcome of an analysis but also demonstrated how results are interactively influenced by the choice of analysis and the number of actors in the stimulus set.

Limitations and practical considerations

The shooter-bias multiverse analysis was limited by several factors. First, only 19 studies were included. Stronger conclusions could have been drawn if a larger number of data sets were available. In addition, there was an issue of range restriction in the number of targets these studies included, and this may have limited the ability of the analysis to detect patterns across studies because it meant that most studies were underpowered. Moreover, an examination of Table 1 indicates that sample size, number of targets, and response window were also somewhat confounded with each other, although this pattern of confounding cannot explain most of the current findings. Finally, results for certain study-analysis combinations are missing because analyses did not reach convergence; if those unknown results differed systematically from the known results in some way, it is possible that excluding them would alter the multiverse analysis's conclusions. These limitations of the current example illustrate some practical considerations that should inform the use of the multiverse-of-methods approach and that largely stem from potential deficiencies of available data.

Many of the limitations of the shooting-bias example are related to a lack of data representing certain designs or combinations of designs. This is likely to be a common problem for projects examining a multiverse of methods. Often, the entire multiverse of methods—that is, every possible combination of decision options—will not be available in the published literature. Here is a simple example: Suppose there are two data-collection decisions that are of interest—whether to use rats or mice and whether to place electrodes at Site A or Site B. It may be the case that past research has included both rat studies and mouse studies, as well as both Site A studies and Site B studies, but all of the rat studies have placed electrodes at Site A. Multiverse results would be informative as to the effect of electrode placement for mice and the effect of species when measuring at Site A but would not indicate whether the effect of electrode placement might be different for rats than for mice. At other times, not every reasonable decision option will be available in the literature even for a single variable, as in the current example in which no study used more than 50 targets. Such situations are particularly likely when examining small bodies of research. When drawing conclusions, researchers should therefore take care to consider how results are limited by unavailable data.

Moreover, when results from the multiverse are themselves missing, as in the current example when some analyses met with convergence issues, some effort should be made to determine whether missingness is systematic and/or has influenced conclusions. The shooter data may be illustrative here. In the current data, convergence issues typically occur in the error analyses and are more common among studies with longer response windows—which are also likely to have few errors and therefore less variability in the response variable. It so happens that these studies are also likely to have small sample sizes and small numbers of targets. All three of these are factors limiting the data's capacity to support estimating complex models. This provides some explanation for the issues with convergence but also means that results are not missing at random. A potential concern is that conclusions may be limited to the subgroup of results that converged. Considering other patterns in the data may shed light on this. For study-level variables such as response window, it is possible to partially screen for the convergence dependence of results by comparing results across studies for only the types of analyses that always converged (i.e., the rows in the figures labeled "ANOVA," "Fixed," and "IT"). Conclusions do not change if only these rows are considered.

Another consideration is that in most cases, if a study finds a significant effect for one analysis, it finds a significant effect for every converging analysis between the row labeled "Fixed" and the row labeled "IRP + IT." As a thought experiment, we can fill in the nonconverging cells with hypothetical results consistent with this pattern; doing so leads to the same conclusions as does examining the results with those cells excluded.[9] Excluding the cells that did not converge, therefore, would affect the conclusions only if these cells had a tendency to violate the overall pattern. There is no theoretical reason to expect this—the study-level variables that are related to nonconvergence are not associated with marked violations of this pattern—but it is technically possible. Thus, it seems unlikely that the cells that did not converge are hiding alternative patterns of results, but it cannot be ruled out as a possibility. Questions such as these must be considered when certain analyses cannot be performed on certain studies.

Finally, it is worth noting that comparing data from multiple studies carries the limitation that conditions and samples may vary across studies in unknown ways. Two studies conducted by different labs at different times may differ in a host of nuisance variables that will confound whatever difference between the two studies is actually of interest. Conclusions from multiverse analyses will be most robust when multiple studies are available per cell and when the studies in any two cells do not vary systematically in any nuisance variable (e.g., lab of origin).

Types of methodological ambiguity

The primary purpose of a multiverse analysis is to clarify how methodological and analytic decisions affect results. This can be important for either of two reasons depending on the decision in question.

The first of these reasons is that for some of the choices researchers face, there is reason to believe that one particular option is the "correct"—or at least better—way to do things. For example, if researchers vary in terms of the statistical analyses they use to address a particular question, it is often the case that one of these analyses is more appropriate than the other options. Or a particular data-collection method may generate less measurement noise than an alternative. These are the researcher decisions for which it would be useful to establish a consensus regarding best practice. In these cases, variation in practice is of concern simply because not all researchers are conducting their studies with optimal methods.

The second reason applies to decisions for which the question of which option is the "best" is more complicated, as reasonable arguments could be made for multiple choices. In these cases, it may not be possible to answer the question of which option is best. Instead, the important question is whether the different options are likely to lead to different study conclusions. If they are, then there is a risk that researchers may try out multiple methods or analyses, selectively report results from the ones that produced significant effects, and ignore those that "did not work," contributing to an overrepresentation of Type I errors in the published literature. Analytic options that enable such a process are often referred to as *researcher degrees of freedom* (Simmons, Nelson, & Simonsohn, 2011). However, the concept is applicable to methodological options as well because ambiguity about which methods to use can contribute to publication bias in a process directly analogous to the process by which analytic options inflate Type I error rates. For example, a researcher may run two or three shooter studies, trying out a different response window each time, until one of them shows a significant effect of the researcher's hypothesized intervention or moderator. After obtaining a significant effect, the researcher is likely to report only this study, with hindsight-fueled confidence that it had obviously been the correct choice all along. Such practices can lead a researcher to inadvertently publish spurious effects. Developing a consensus on how to make such methodological decisions is therefore desirable to remove ambiguity about which method is best and reduce the frequency of these situations.

Applications of the multiverse-of-methods analysis

A multiverse-of-methods analysis can be used to confront either of these sources of methodological variation. When one method is clearly superior to another, that superiority can usually be established through statistical reasoning or experimentation, without the use of a multiverse analysis. However, the task of persuading researchers that this difference is important enough to merit changing their methods can sometimes be difficult. As controversy over reproducible methods demonstrates, there is variation in the extent to which researchers adopt best-practice methods (John, Loewenstein, & Prelec, 2012). If the inferior methods are more familiar, or if implementing the superior methods would require obtaining new knowledge or equipment, adopting the superior methods could lead researchers to incur some short-term costs in effort, time, or money. In these situations, motivated reasoning can lead researchers to discount arguments that they should make such methodological changes. A multiverse analysis, by providing a concrete demonstration of the consequences of a methodological choice, can be a useful persuasive tool in these methodological controversies.

The multiverse-of-methods analysis is also useful when multiple methodological alternatives exist but it is unclear whether any one of them is "better" than the others. In these cases, a multiverse analysis can clarify which of the alternatives differ from one another in terms of the typical results and whether this depends on other factors—as an example, consider how the influence of response window on shooter-study conclusions depended on whether errors or reaction time were used as the dependent variable in the example above. Understanding such patterns can shed light on whether and why some methods might be preferable overall or preferable for certain research questions or designs. This can provide direction for subsequent research investigating the causes of the differences observed in the multiverse. Multiverse results may provide this direction by identifying which alternatives differ from each other, especially when there are several alternatives, and/or by suggesting hypotheses for why these differences exist (e.g., through their interactions with other factors in the multiverse). By guiding research on the implications of various methods in this way, the multiverse-of-methods approach can assist with identifying best practices, improving research efficiency and reducing ambiguity that researchers may inadvertently exploit to obtain significant results.

Moreover, when a multiverse analysis reveals that study conclusions do not vary across a set of methodological alternatives, this provides evidence that these alternatives may be equally valid options. That said, it should be noted that failing to find evidence for a difference is not the same as finding evidence for no difference, and further research will always be advisable to confirm a lack of difference among studies, particularly if the multiverse included a small sample size of studies. However, such a finding is at least suggestive of no difference among methods, which is useful when evaluating past studies that used differing methods (e.g., when comparing two studies' results or when making decisions about which studies to include in a meta-analysis). Finding no evidence for a difference among methods also suggests that future researchers may choose among these methods on the basis of practical considerations such as efficiency or expense without concern that they are compromising the quality of their studies.

Conclusion

Through this extension of the multiverse analysis, researchers can address ambiguity around methodological decisions that can obscure best practice and even inflate Type I error rates in the published literature. It has a role both when best practice is known but not universally implemented as well as when multiple methods are used across studies but it is unclear whether they produce differing results. In short, the multiverse-of-methods analysis is useful in a variety of ways: as a persuasive tool, as a tool for narrowing down questions about how and why methodological alternatives produce different results, and as a source of information about how to evaluate past studies and design future studies.

Transparency

Action Editor: Laura A. King
Editor: Laura A. King
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iD

Jenna A. Harder https://orcid.org/0000-0003-0751-8438

Acknowledgments

Many thanks to Deborah Kashy, Joseph Cesario, Richard Lucas, and Victor Keller for their comments on drafts of this article. Thanks also to Joseph Cesario, Joshua Correll, Jared Kenworthy, Debbie Ma, Sang Hee Park, Jessica Sim, and Audrey Snowden for sharing their data.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/1745691620917678

Notes

[1] Other studies have analyzed error and response-time data in combination using the drift-diffusion model (e.g., Pleskac, Cesario, & Johnson, 2018). However, this is a less common approach and addresses process-level questions rather than behavioral racial bias; the current example therefore does not consider it as one of the analytic options to be compared.

[2] Shooter studies have also used various models designed to provide information about the cognitive processes underlying the behavioral data, such as signal detection theory (e.g., Correll, Wittenbrink, Park, Judd, & Goyle, 2011) and the drift-diffusion model (see previous note). However, the discussion of these models is beyond the scope of this article.

[3] For a helpful review of multilevel modeling and the importance of accounting for stimulus as a grouping factor, see Judd, Westfall, and Kenny (2012). For a more in-depth study of multilevel modeling, see Hox, Moerbeek, and van de Schoot (2018).

[4] The most complete possible random-effects specification would allow intercepts to vary randomly by both participant and target, allow object slopes to vary randomly by both participant and target, allow race slopes to vary by participant, and allow the slope for the Race × Target interaction to vary randomly by participant. However, the current multiverse analysis did not allow more than one slope at a time to vary by participant.

[5] SPSS was used for these analyses because *lme4*, the R package used in the main multiverse analysis, does not calculate statistical significance for random effects.

[6] In both error and reaction-time analyses, coefficients did not appreciably change with increasing analysis complexity.

[7] However, an examination of Race × Object interaction coefficients suggests that using greater numbers of targets may be slightly associated with smaller degrees of bias toward shooting Black targets in error and reaction-time analyses (see the Supplemental Material available online).

[8] It may be of use to readers to know that there are some ways to make convergence issues less likely. One helpful tip is to mean-center continuous variables and effects-code categorical variables. Other strategies are software-specific; tips for achieving convergence in R can be found online ("lme4 convergence," n.d.). However, one common cause of nonconvergence is simply that the random-effects structure that is specified is too complex to be specified with the available data; that is, the model is "overparameterized" (Bates, Kliegl, Vasishth, & Baayen, 2015). Bates et al. (2015) provide a useful explanation of this issue and advice for making decisions about which parameters to discard when simplifying a model.

9. For example, sample size continues to be unrelated to the significance of error analyses: 39% versus 41%. I do not walk through the arithmetic here for the sake of space, but readers can perform this thought experiment for themselves using Figures 2, 4, and 6.

References

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv*. Retrieved from https://arxiv.org/pdf/1506.04967v1.pdf

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*, 1314–1329.

Correll, J., Urland, G. R., & Ito, T. A. (2006). Event-related potentials and the decision to shoot: The role of threat perception and cognitive control. *Journal of Experimental Social Psychology*, *42*, 120–128.

Correll, J., Wittenbrink, B., Park, B., Judd, C. M., & Goyle, A. (2011). Dangerous enough: Moderating racial bias with contextual threat cues. *Journal of Experimental Social Psychology*, *47*, 184–189.

Credé, M., & Phillips, L. A. (2017). Revisiting the power pose effect: How robust are the results reported by Carney, Cuddy, and Yap (2010) to data analytic decisions? *Social Psychological and Personality Science*, *8*, 493–499.

Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Bastian, B., . . . Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *Journal of Personality and Social Psychology*, *114*, 323–341.

Eiler, B. A. (2017). *The behavioral dynamics of shooter bias in virtual reality: The role of race, armed status, and distance on threat perception and shooting dynamics* (Doctoral dissertation). Retrieved from PsycINFO. (Order No. AAI10760 366)

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). New York, NY: Routledge.

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology*, *108*, 187–218. doi:10.1037/a0038557

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.

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

Lme4 convergence warnings: Troubleshooting. (n.d.). Retrieved from https://rstudio-pubs-static.s3.amazonaws.com/33653_57fc7b8e5d484c909b615d8633c01d51.html

Ma, D. S., & Correll, J. (2011). Target prototypicality moderates racial bias in the decision to shoot. *Journal of Experimental Social Psychology*, *47*, 391–396.

Mekawi, Y., & Bresin, K. (2015). Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis. *Journal of Experimental Social Psychology*, *61*, 120–130.

Miller, S. L., Zielaskowski, K., & Plant, E. A. (2012). The basis of shooter biases beyond cultural stereotypes. *Personality and Social Psychology Bulletin*, *38*, 1358–1366.

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event-and interval-contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin*, *27*, 771–785.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass*, *2*, 842–860.

Park, S. H., & Glaser, J. (2011). Implicit motivation to control prejudice and exposure to counterstereotypic instances reduce spontaneous discriminatory behavior. *Korean Journal of Social and Personality Psychology*, *25*, 107–120.

Park, S. H., Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice moderates the effect of cognitive depletion on unintended discrimination. *Social Cognition*, *26*, 401–419.

Park, S. H., & Kim, H. J. (2015). Assumed race moderates spontaneous racial bias in a computer-based police simulation. *Asian Journal of Social Psychology*, *18*, 252–257.

Pleskac, T. J., Cesario, J., & Johnson, D. J. (2018). How race affects evidence accumulation during the decision to shoot. *Psychonomic Bulletin & Review*, *25*, 1301–1330.

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Sim, J. J., Correll, J., & Sadler, M. S. (2013). Understanding police and expert performance when training attenuates (vs. exacerbates) stereotypic bias in the decision to shoot. *Personality and Social Psychology Bulletin*, *39*, 291–304.

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*, 1359–1366.

Snowden, A. K. (2017). *Induced emotions on shoot decisions* (Doctoral dissertation). Retrieved from https://rc.library.uta.edu/uta-ir/bitstream/handle/10106/26759/SNOWDEN-DISSERTATION-2017.pdf

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.